Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 860–869 doi: 10.1111/j.2041-210X.2012.00225.x

Designing studies to detect differences in species occupancy: power analysis under imperfect detection

Gurutzeta Guillera-Arroita* and José J. Lahoz-Monfort

National Centre for Statistical Ecology, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK

Summary

1. Studies aimed at estimating species site occupancy while accounting for imperfect detection are common in ecology and conservation. Often there is interest in assessing whether occupancy differs between two samples, for example, two points in time, areas or habitats. To ensure that meaningful results are obtained in such studies, attention has to be paid to their design, and power analysis is a useful means to accomplish this.

2. We provide tools for conducting power analysis in studies aimed at detecting occupancy differences under imperfect detection and explore associated design trade-offs. We derive a formula in closed form that conveniently allows determining the sample size required to detect a difference in occupancy with a given power. Because this formula is based on asymptotic approximations, we use simulations to assess its performance, at the same time comparing that of different significance tests.

3. We show that the closed-formula performs well in a wide range of scenarios, providing a useful lower sample size bound. For the simulated scenarios, a Wald test on the probability scale was the most powerful test among those evaluated.

4. We found that choosing the number of repeat visits based on existing recommendations for single-season studies will often be a good approach in terms of minimizing the effort required to achieve a given power.

5. We demonstrate that our results and discussion are applicable regardless of whether independence or Markovian dependence is assumed in the occupancy status of sites between seasons, and illustrate their utility when designing to detect a trend in multiple-season studies.

6. Assessing differences in species occupancy is relevant in many ecological and conservation applications. For the outcome of these monitoring efforts to be meaningful and so to avoid wasting the often limited resources, survey design has to be carefully addressed to ensure that the relevant differences can be indeed detected and that this is achieved in the most efficient way. Here, we provide guidance and tools of immediate practical use for the design of such studies, including code to conduct power analysis.

Key-words: hypothesis testing, likelihood-ratio test, multiple-season occupancy, sample size, survey design, Wald test

Introduction

Occupancy, defined as the proportion of sites occupied by a species, is a state variable of interest in various areas of ecology (MacKenzie *et al.* 2006). In most cases, species detection is imperfect, which can lead to the incorrect classification of occupied sites as empty. If imperfect detection is not accounted for, bias is induced in the occupancy estimator. To tackle this problem, MacKenzie *et al.* (2002) and Tyre *et al.* (2003) proposed the joint modelling of occupancy and detection probabilities based on data resulting from a sampling protocol in which discrete replicate surveys are carried out at each sampling site, a modelling framework that has become widely used by ecologists.

To ensure that occupancy studies provide meaningful results and that, therefore, valuable monitoring resources are not wasted, it is critical to pay attention to survey design (Yoccoz, Nichols & Boulinier 2001; Legg & Nagy 2006). Unfortunately, enough care is not always devoted to this important stage, and providing simple tools to facilitate the process can help

*Correspondence author. E-mail: gg63@kent.ac.uk

promote its rightful consideration. For instance, in the context of single-season occupancy studies, tables have been developed to assist in the allocation of survey effort between number of sites and replicate visits (MacKenzie & Royle 2005; Guillera-Arroita, Ridout & Morgan 2010). Sample size can then be determined by setting a target variance for the occupancy estimator, either choosing the minimum number of sites to achieve this target or as many as allowed by the available effort (and checking whether the variance target is met).

Rather than making inference about species occupancy at a given point in time, area or habitat type, there is often interest in assessing whether there are differences in occupancy between two samples. Occupancy has been proposed as a useful state variable for various large-scale monitoring programmes (MacKenzie *et al.* 2006, pp 41–44), with occupancy declines being of particular interest when dealing with species of conservation concern and increases when tracking the spread of invasives. In IUCN Red List assessments, estimated occupancy declines are used as part of criteria A and B, to reflect declines of population size and geographical range (IUCN 2001). In active adaptive management (McCarthy & Possingham 2007), the difference in species occupancy before and after the experimental intervention can be a state variable of interest. The focus might also be in assessing occupancy differences between two geographical areas or habitat types. In studies aimed at detecting occupancy differences, the criterion for sample size selection can be expressed in terms of their desired power, that is, the probability that the study will detect a significant difference, given that the true difference is of a given size (Cohen 1988).

The concept of power analysis is closely related to that of null hypothesis significance testing (NHST). As an inferential procedure, NHST has long received criticism (for a comprehensive review see Nickerson 2000), among other reasons because it lends itself to confusion regarding statistical vs. scientific significance. We do not extend here on this issue as there is a wealth of papers addressing it, including many in the context of ecological studies (e.g. Yoccoz 1991; Cherry 1998; Johnson 1999; Di Stefano 2004). A common understanding is that it is better to focus on providing an estimation of the magnitude of the effect in the form of confidence intervals, as this conveys more information than the outcome of a significance test. However, despite the NHST controversy, prospective power analysis is widely recognized as a useful tool for study design (Cohen 1990; Thomas & Juanes 1996; Steidl, Hayes & Schauber 1997; Johnson 1999; Di Stefano 2001; Legg & Nagy 2006). In fact, a particularly beneficial aspect of power analysis is that it requires an explicit consideration of what constitutes a biologically significant result, allowing us to determine whether a given design renders our study a good chance of producing statistically significant results when the actual effect size (occupancy difference in our case) is biologically significant.

While simulations provide a tool for power analysis, they can be time-consuming. Closed formulae can sometimes be derived to determine more easily the sample size required to achieve a given power. The development and performance evaluation of such formulae for a test comparing two independent binomial proportions have received a lot of attention in the literature (e.g. Cochran & Cox 1957, p. 27; Fleiss 1973, p. 30; Casagrande, Pike & Smith 1978; Walters 1979; Fleiss, Tytun & Ury 1980; Ury & Fleiss 1980; Dobson & Gebski 1986; Gordon & Watson 1996; Vorburger & Munoz 2006). These formulae are routinely used in different areas, such as the design of clinical trials (Donner 1984). However, as they assume that the outcome of the experiment, whether success or failure, is always observed without error, they are not applicable for occupancy studies, except for the unusual case in which species detection is perfect or enough replicate surveys are carried out to ensure its detection is practically certain.

To our knowledge, sample size formulae for models that account for imperfect detection when comparing binomial proportions have not been proposed or evaluated to date. In this study, we address this problem. We discuss how to design studies to detect a difference in occupancy between two samples with a given power when species detection is imperfect, and we present tools to accomplish this. We provide an approximate expression to calculate power and derive a closed-formula that conveniently allows the number of sites that need to be sampled to be determined with just a few simple calculations, while accounting for species detectability. Using this expression, we examine how the power of a study changes depending on the allocation of survey effort between number of sites and number of replicate visits and thus revisit the issue of optimal replication that had previously been addressed from the point of view of minimizing the variance of the occupancy estimator in single-season studies (MacKenzie & Royle 2005; Bailey *et al.* 2007; Guillera-Arroita, Ridout & Morgan 2010). As the derived sample size formula involves asymptotic (i.e. large sample) approximations, its performance needs to be assessed, as this is essential to understand its applicability. For this, we run Monte Carlo simulations and check how the resulting sample sizes compare to those indicated by the formula, at the same time also evaluating the performance of various significance tests. In the context of studies that assess occupancy changes in time, we demonstrate that the results and discussion in the paper are applicable regardless of whether independence or Markovian dependence is assumed in the occupancy status of sites between seasons (MacKenzie *et al.* 2006, pp. 186–212), and illustrate their utility when designing to detect a trend in multiple-season studies. Finally, we provide R code for conducting power analysis, both based on the formula and via simulations (Appendix S3).

Power analysis calculations

Statistical tests assess evidence to reject the null hypothesis of no effect. For instance, when comparing occupancy estimates from two seasons, the interest often lies in assessing whether occupancy has changed in this period. Because of their probabilistic nature, statistical tests always have the possibility to (i) detect an effect when there is no such effect (false positive or Type I error) and (ii) not detect an existing effect of a given size (false negative or Type II error). The probability of Type I error or 'significance level' is usually denoted by α and the probability of Type II error by β. The power of a statistical test

is the probability of detecting an effect, given there is an effect of a given size (i.e. 1– β, Appendix S1-Fig. A1). A significance level commonly chosen in statistical tests is 0·05, and in terms of power, levels around 0·8 are often used. Throughout this study, we use $\alpha = 0{\cdot}05$ for illustration, but, in practice, α and β should reflect the relative seriousness of Type I and II errors (see Discussion). The concept of power is intrinsically related to the concept of 'effect size', that is, how large the effect that we wish to detect is. For a given sample size and significance level, the larger the effect size under consideration, the greater the power of the test.

MODELLING OCCUPANCY

The single-season occupancy modelling framework is based on a sampling protocol in which S sites are repeatedly surveyed recording whether the species is detected. The model assumes that each site has a probability of being occupied and that the detection process at occupied sites is a series of independent Bernoulli trials. Assuming constant occupancy ψ and detectability p, and a standard sampling design with K replicate surveys per site, the likelihood for the model is

$$L(\psi, p|h) = \prod_{i=1}^{S} \left\{ \psi p^{d_i} (1-p)^{K-d_i} + (1-\psi) I(d_i = 0) \right\}, \quad \text{eqn 1}$$

where d_i is the number of detections at site i, and I (.) takes value one when the expression in brackets is true and zero otherwise. There are two explanations for sites without detections: either the site was empty or it was occupied and the species was missed in all surveys. The model assumes that there are no false positives, which are generally a much lesser problem (but see Miller *et al.* 2011).

For this model, the expressions for the maximum-likelihood estimators (MLEs) and their asymptotic variance–covariance matrix are known (Guillera-Arroita, Ridout & Morgan 2010). In particular, the asymptotic variance of the occupancy estimator is

$$\sigma^2 = \frac{\psi}{S} \left\{ (1-\psi) + \frac{1-p^*}{p^* - Kp(1-p)^{K-1}} \right\}, \quad \text{eqn 2}$$

where $p^* = 1 - (1-p)^K$ is the probability of not missing the species at an occupied site (i.e. detecting it in at least one of the K surveys). Equation 2 has the form of the variance of a binomial proportion with an extra term $F = (1-p^*)/\{p^* - Kp(1-p)^{K-1}\}$ introduced by imperfect detection (Appendix S1-Fig. A2), which tends to zero as p^* tends to 1.

When the interest is in comparing occupancy between two samples for which the occupancy status of sites can be considered independent, the analysis can be carried out by applying the model in eqn 1 separately to each of the data sets. Two samples can be considered independent if they consist of different sampling sites, for instance when comparing occupancy between two geographical locations or habitat types. Studies assessing changes between two points in time may also sample different sites, although often the same sites are sampled in both seasons. Even so, the assumption of independence can still be valid, for example, when dealing with species that display a low degree of site fidelity or when the time elapsed between seasons is sufficiently long, so that the observed changes can effectively be considered random (MacKenzie *et al.* 2006, p. 206).

[Correction added on 14 Sept after first online publication: the wording of the following paragraph has been altered.] Hereafter we assume independence in the occupancy status of sites between samples. For those cases with dependence between seasons, data can be analysed with a model that explicitly describes the process underlying occupancy dynamics as a first-order Markov chain (MacKenzie *et al.* 2003). In Appendix S2, we derive the MLE expressions and variance-covariance matrix for this model and discuss the practical implications of the lack of independence for study design. Note that assuming independence provides a conservative design in the most common dependence scenarios.

TESTING FOR SIGNIFICANCE IN OCCUPANCY DIFFERENCES

Various approaches can be used for testing the null hypothesis of no difference in occupancy between two samples (ψ_1 and ψ_2). One possibility is to determine significance based on a z-test. Maximum-likelihood theory shows that, asymptotically, the occupancy *estimator* is unbiased and normally distributed $\hat{\psi} \sim N(\psi, \sigma^2)$. Therefore, the random variable $\hat{D} = \hat{\psi}_1 - \hat{\psi}_2$is normally distributed $\hat{D} \sim N(\psi_1 - \psi_2, \sigma_1^2 + \sigma_2^2)$. Under the null hypothesis $\hat{D} \sim N(0, \sigma_1^2 + \sigma_2^2)$ so, for a significance level α, the critical region for a two-tailed test is bounded by $\pm z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2}$, where $z_{\alpha/2}$ is the upper 100α/2-percentage point for the standard normal distribution. Consequently, a difference would be considered significant if $\left|\hat{\psi}_1 - \hat{\psi}_2\right| / \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} > z_{\alpha/2}$, where $\hat{\psi}_i$ are the maximum-likelihood occupancy *estimates* and $\hat{\sigma}_i$ their estimated standard errors (SE). This is in fact a Wald test (Morgan 2008, p. 101) and is equivalent to assessing whether the confidence interval (CI) for the estimated difference includes zero. In practice, the likelihood maximization is usually performed via a logistic reparameterization. The Wald test could also be performed on this scale, considering a difference significant if $|\hat{\beta}_1 - \hat{\beta}_2| / \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2 + \hat{\sigma}_{\hat{\beta}_2}^2} > z_{\alpha/2}$, where $\hat{\beta}_i = \text{logit}\left(\hat{\psi}_i\right)$ and $\hat{\sigma}_{\beta i}$ their corresponding SEs. This is equivalent to assessing whether $\hat{\beta}_2'$ is significantly different from zero when occupancy in the second sample is instead parameterized as $\beta_1 + \beta_2' = \text{logit}(\psi_2)$.

Another possibility is to carry out a likelihood-ratio test (Morgan 2008, p. 80), which compares the fit of two models, where one (the null) is a special case of the other (the alternative). The null model here has a common parameter for occupancy across both samples ($\psi_1 = \psi_2$), while the alternative allows for different occupancy parameters. A difference would be considered significant at the α level if $-2L_0 + 2L_A > \chi^2_{\alpha:1}$, where L_0 and L_A are the maximum log-likelihood values for the null and alternative models, respectively, and $\chi^2_{\alpha:1}$ is the upper 100α-percentage point for the chi-square distribution

with one degree of freedom. Both the likelihood-ratio and Wald tests are based on asymptotic approximations and are asymptotically equivalent under the null hypothesis.

A FORMULA TO ASSESS POWER TO DETECT A DIFFERENCE IN OCCUPANCY

A formula to assess the power to detect a difference in occupancy that would be achieved under a given study design and underlying probabilities of occupancy and detection can be derived considering again the properties of the estimators. We assume here a standard sampling design with K replicate surveys carried out at S sampling sites, and constant probabilities of occupancy and detection. Let ψ_1 and ψ_2 be the true underlying occupancy probabilities in the two samples. Given that the critical region for a two-tailed test is bounded by $\pm z_{\alpha/2}\sqrt{\sigma_1^2+\sigma_2^2}$, the probability of observing an occupancy difference that falls within the critical region (i.e. power) is

$$G = 1-\beta = \left\{1-\Phi\left(\frac{z_{\alpha/2}\sqrt{\sigma_1^2+\sigma_2^2}-(\psi_1-\psi_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right\} + \Phi\left(\frac{-z_{\alpha/2}\sqrt{\sigma_1^2+\sigma_2^2}-(\psi_1-\psi_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \quad \text{eqn 3}$$

where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution, and $\sigma_i^2 = \psi_i(1-\psi_i+F_i)/S_i$

Let R be the proportional difference in occupancy, so that $\psi_2 = \psi_1\ (1-R)$, with $R > 0$ representing a decline and $R < 0$ an increase. Note that $R \in [(\psi_1 - 1)/\psi_1, 1]$ to ensure $\psi_2 \in [0, 1]$. The plot of G as a function of effect size (R here) is known as the 'power curve' of the test (Fig. 1). All power curves pass through $(0, \alpha)$ because an effect of magnitude 0 corresponds to the null hypothesis which by definition is rejected with probability α. As R increases, the probability of rejecting the null hypothesis increases. For a given R, power increases as the number of sampling sites increases (Fig. 1a). Power also increases with the number of replicate surveys (Fig. 1b), approaching the power expected for a binomial experiment with perfect detection as p^* tends to one (but note that here increasing replication implies an increase in total effort; we discuss later how power changes depending on how a fixed effort is allocated into sites and replicate surveys). A similar behaviour takes place for increases in detection probability, with power saturating as p tends to one (Fig. 1c). The larger the initial occupancy probability ψ_1, the larger the power to detect a given proportional difference R, as this translates into a larger absolute occupancy difference (Fig. 1d). Power increases with α, but at the expense of accepting more false positives.

A FORMULA TO DETERMINE SAMPLE SIZE

Equation 3 can be solved numerically to determine the number of sites that need to be surveyed to achieve a given power.

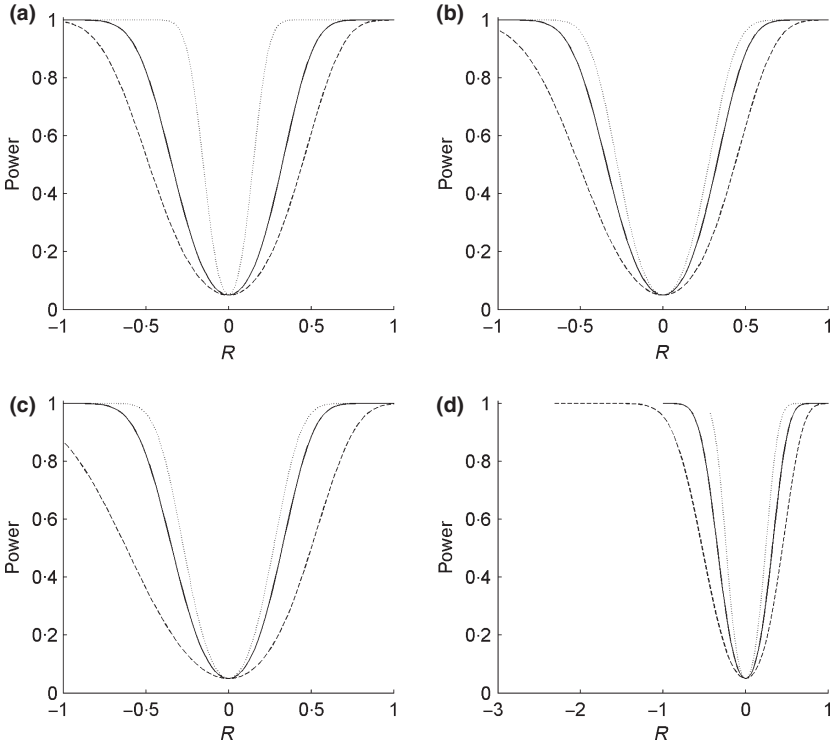

Fig. 1. Power curves ($\alpha = 0{\cdot}05$). The solid line represents a reference case with $\psi_1 = 0{\cdot}5$ and $p = 0{\cdot}5$, $K = 3$, $S = 100$ for both survey periods. In each panel, one of these parameters changes: (a) $S = 50$ (dash), $S = 500$ (dot); (b) $K = 2$ (dash), $K = 6$ (dot); (c) $p = 0{\cdot}3$ (dash), $p = 1{\cdot}0$ (dot); (d) $\psi_1 = 0{\cdot}3$ (dash), $\psi_1 = 0{\cdot}7$ (dot).

However, an approximation is possible that gives a convenient expression in closed form. Without loss of generality, we assume that $\psi_1 - \psi_2 > 0$. In this case, the second term in eqn 3 can be considered negligible as it represents the probability of detecting an increase when there is a decline, which will be small and corresponds to cases in which an incorrect inference about the sign of the occupancy difference would have been made. The power to detect can be written now as

$$G = 1 - \beta = 1 - \Phi\left(\frac{z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2} - (\psi_1 - \psi_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \qquad \text{eqn 4}$$

Assuming that the same number of sites are to be surveyed in both occasions, and considering that by definition $1 - \beta = \Phi(z_\beta)$, and by symmetry $\Phi(x) = 1 - \Phi(-x)$, the number of sites to be surveyed to achieve a given power can be derived as a function of the significance level and effect size, given ψ_1, p_1, p_2, K_1 and K_2

$$S = \left(\frac{z_{\alpha/2}\sqrt{f_1 + f_2} + z_\beta\sqrt{f_1 + f_2}}{\psi_1 - \psi_2}\right)^2 = (f_1 + f_2)\left(\frac{z_{\alpha/2} + z_\beta}{\psi_1 - \psi_2}\right)^2, \qquad \text{eqn5}$$

where $f_i = \psi_i\ (1 - \psi_i + F_i)$. Equation 5 assumes a null hypothesis of no difference, but can be modified if the focus is to evaluate whether the difference is greater than a threshold D_0 (positive for declines and negative for increases), by changing the denominator to $\psi_1 - \psi_2 - D_0$ and using a one-tailed test.

POWER ANALYSIS AND OPTIMAL DESIGN

It is important to realize that there is a trade-off in how the available resources are allocated into number of sites S and amount of replication K, as this choice affects the power to detect a difference in occupancy. As the number of repeat visits in a sample increases, the precision of the occupancy estimator improves. However, when this is performed at the expense of reducing the number of sampling sites, it creates a trade-off leading to an optimal number of repeat visits which depends on the characteristics of the species. Figure 2 explores this trade-off assuming that the cost of all individual surveys is the same, and thus, defining as a cost function, the total survey effort, $E = S \cdot K$. For each scenario, there is a value of K, given by the minimum of each curve, which provides an optimum survey design in terms of total survey effort required to achieve a given power.

It is also worth investigating the impact of designing a survey away from the optimum, a situation that may arise, for example, because of logistic constraints. These plots can be used to evaluate the extra survey effort required to maintain the same power if the survey design strays from the optimum. As detectability decreases, the optimum K increases and so does the total effort. Note that with low detectability and mid-high values of occupancy, reducing K has a much greater

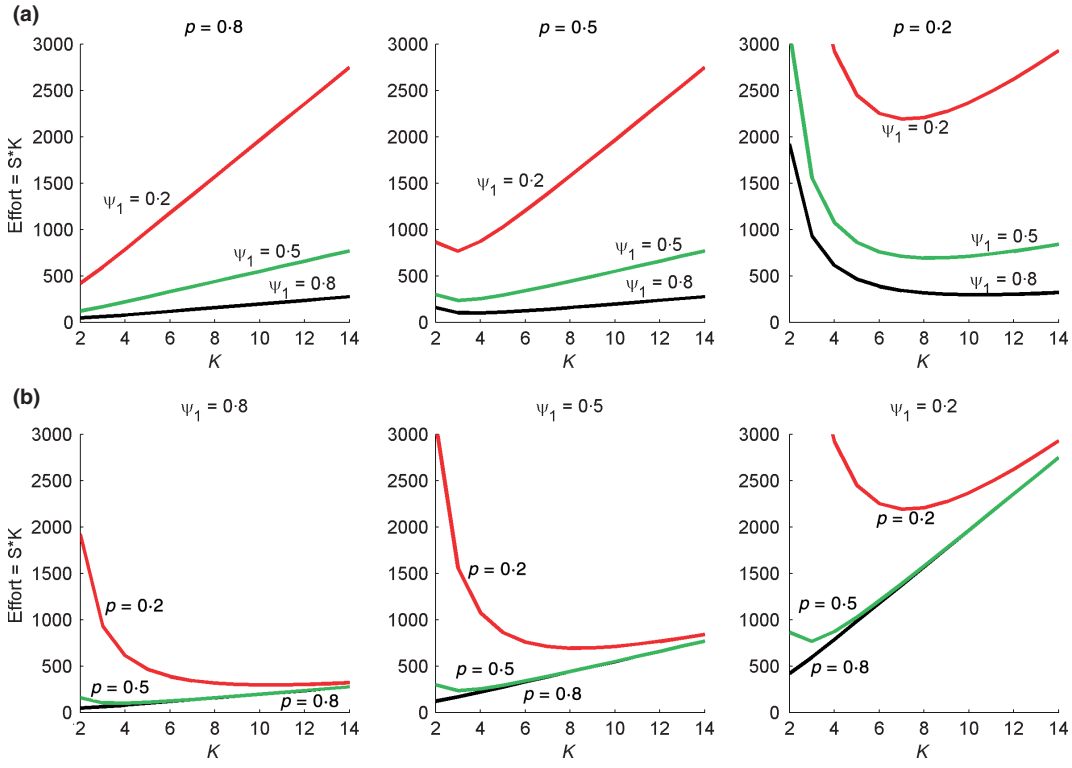

Fig. 2. Minimum survey effort ($E = S \cdot K$) to achieve power 0·8 when the occupancy decline is $R = 0{\cdot}5$, for varying K and different scenarios of initial occupancy ψ_1 and detectability p ($\alpha = 0{\cdot}05$). In both seasons, p and K are kept constant. Top panels display cases of constant p and lower panels constant ψ_1.

impact than increasing it. When p is high (and, therefore, imperfect detection is less of an issue), the design requiring minimum effort is always $K = 2$, and the curves for different values of ψ_1 are basically straight lines. Any effort invested in extra replication ($K > 2$) is 'wasted', and the number of sites required for the target power remains constant. Actually, all curves tend asymptotically to the straight line given by the value of S that fulfils the power requirements assuming perfect detection: given enough replication, the species is never missed in sites where it is present, and the variance of the occupancy estimator no longer depends on detectability. This asymptotic behaviour can be seen clearly in Fig. 2b.

As mentioned in the introduction, survey design guidelines exist for selecting a K that minimizes the variance of the occupancy estimator. Interestingly, for the scenarios in Fig. 2, where detectability is kept constant in both samples, we found that the value of K that follows this design criterion for the first survey season also corresponded to that achieving the target power with minimum total survey effort. There was only some slight discrepancy in more extreme cases of very low detectability together with high initial occupancy, as then the optimal K to minimize the variance of ψ_1 differs more from that required to minimize the variance of ψ_2. In these cases, the optimal K in terms of power was slightly lower, as we were assessing a decline, and thus the optimal K to minimize the variance of ψ_2 would be lower than that of ψ_1.

Power analysis simulations

SIMULATION PROCEDURE

Equation 5 conveniently allows conducting a power analysis with just a few simple calculations, but power can also be evaluated via simulation. This involves generating and analysing two detection histories with the assumed values for occupancy and detection probability, and a given study design, and assessing whether a significant occupancy difference is detected. By repeating the simulations a large enough number of times, power can be computed as the proportion of simulations in which a significant difference is detected.

As eqn 5 is based on asymptotic approximations, we ran simulations to assess its performance. We compared the survey effort required to achieve a given power to detect an occupancy decline as indicated by both approaches under various scenarios. We explored different values of initial occupancy ($\psi_1 = 0{\cdot}2, 0{\cdot}5, 0{\cdot}8$), detection probability ($p = 0{\cdot}2, 0{\cdot}5, 0{\cdot}8$) and effect size ($R = 0{\cdot}3, 0{\cdot}5$) under a range of replication levels ($K = 2,\ldots,6$), assuming independence between seasons. We used a significance level $\alpha = 0{\cdot}05$ and simulated scenarios with increasing levels of total survey effort in 10% steps, starting from the survey effort indicated by the formula to achieve power = 0·7, and stopping when the simulations achieved power > 0·9. We ran 5000 simulations per scenario, which should provide sufficiently precise power estimates, (for power = 0·5, the most demanding case, SE = 0·007). In each of the simulations, we obtained maximum-likelihood estimates in MATLAB (The MathWorks Inc., Natick, MA, USA), using the optimization function *fminsearch* on the logistic scale. SEs were derived with the function *mlecov,* which returns an approximation to the asymptotic variance–covariance matrix, and transformed to the probability scale using the delta method. For each simulation, we assessed significance according to three methods: Wald tests on the probability and logistic scales, and a likelihood-ratio test. We also ran simulations to verify the size of the tests (i.e. the attained probability of Type I error, which is the power when $R = 0$) for the scenarios described above with the design suggested by the formula for $R = 0{\cdot}5$ and power = 0·8.

SIMULATION RESULTS

Figure 3 compares the simulation results with curves obtained using eqn 5, for $R = 0{\cdot}5$ and power = 0·8. In all cases, the sample size determined by the closed-formula was similar to that indicated by the simulations, which tended to suggest somewhat higher sampling effort. The simulation results based on the likelihood-ratio test indicated that a greater sampling effort was required compared to the Wald test on the probability scale. The results from the Wald test on the logistic scale were very similar to those from the Wald test on the probability scale when the probability of occupancy was not high. When $\psi_1 = 0{\cdot}8$, the Wald test on the logistic scale appeared to have considerably lower power, and therefore, greater sampling efforts were required according to this test. The size of the Wald test on the probability scale and the likelihood-ratio test was reasonably close to the target (Appendix S1-Fig. A3). The sample size simulation results for $R = 0{\cdot}3$ (Appendix S1-Fig. A4) showed more agreement between formula and simulations, which was expected as detecting a smaller effect implies larger sample sizes, and thus less discrepancy with the large sample approximations.

Discussion

The ultimate target of occupancy studies is often to detect potential occupancy differences. The interest might be in assessing differences temporally (e.g. has occupancy changed since the last survey?) or spatially (e.g. is occupancy different in these two areas or habitat types?). These types of question can be relevant in various applications, from theoretical ecological studies to more applied impact assessments. In fact, our work was originally motivated by the need to provide design input for the National Amphibian and Reptile Recording Scheme in the UK (D. Sewell, unpublished data). Understanding sample size requirements and design trade-offs is essential to ensure that meaningful information is fed back to decision-making and management in the most efficient way and that, therefore, the resources invested in the study are not wasted (Legg & Nagy 2006). The results of power analysis may flag whether a study is not worth doing, if the difference considered meaningful cannot be detected with reasonable probability using the affordable sample size.

We derived and evaluated a formula that, with just a few simple calculations, conveniently allows determination of the

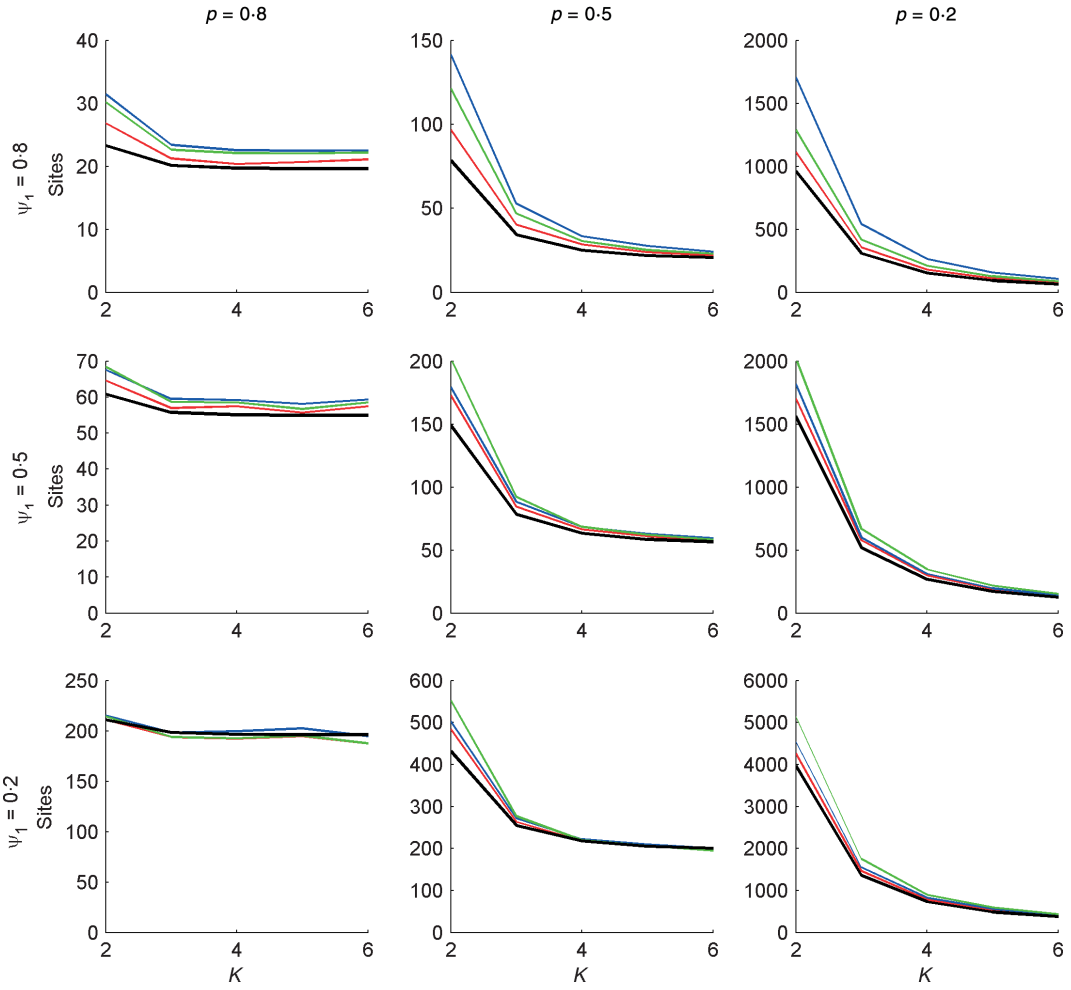

Fig. 3. Number of sampling sites to achieve power 0·8 to detect an occupancy decline $R = 0{\cdot}5$ as indicated by the formula (thick line) and simulations (red: Wald test on probability scale; blue: Wald test on logistic scale; green: likelihood-ratio test), for varying number of repeat surveys K and different scenarios of initial occupancy ψ_1 and detection probability p ($\alpha = 0{\cdot}05$). A varying scale is used for the y-axis to allow detailed comparison but note that, given ψ_1, many more sites need to be visited to achieve a given power when p is low and that, for any level of p, the number of sites tends asymptotically to the same value as K increases.

sample size required to achieve a given power to detect a difference in occupancy under imperfect detection. Having a closed-formula facilitates the process of study design and can, therefore, encourage more attention to be paid to this important but often disregarded step (Yoccoz, Nichols & Boulinier 2001; Legg & Nagy 2006). It can also aid the incorporation of occupancy as a monitoring tool within the broader frameworks of decision theory (Shea & Possingham 2000) and active adaptive management (McCarthy & Possingham 2007), where having an analytical expression enables mathematical optimizations to be performed. It should, however, be kept in mind that the formula is based on approximations and its outcome represents a lower bound. In some situations, more effort might be needed in the study, as illustrated in the results of our simulations. This highlights the benefit of using simulations to assess and potentially refine the initial design (Guillera-Arroita, Ridout & Morgan 2010). The discrepancy between the formula and the simulation results is because of the underlying assumptions not being closely met for some sample sizes. For instance, for the case $\psi_1 = 0{\cdot}8$, $p = 0{\cdot}5$ and $K = 2$ in Fig. 3, the asymptotic variance for the occupancy estimator for $\psi_2 = 0{\cdot}4$ when 78 sites are surveyed is 0·0081 (eqn 2), while the true variance according to simulations is about 50% higher (0·0123). This underestimation explains, at least partly, why the sample size required according to the formula (78 sites) was smaller than that suggested by the simulations (96). The formula expected more precise occupancy estimates so that it would be easier to detect differences among them.

Regarding the performance of significance tests, in our simulations, the Wald test on the probability scale performed better than the likelihood-ratio test (higher power while having a size close to the nominal significance level), which suggests it is a good choice for this type of study. The Wald test is not invariant under a reparameterization and can produce different outcomes depending on how it is performed. Indeed, in our results, the Wald test on the logit scale showed decreased power compared to the other tests when initial occupancy was high. It has been shown that the Wald test can produce misleading results when working with discrete probability distri-

butions under certain parameterizations (Vaeth 1985), including that the test can display aberrant behaviour when testing the equality of two proportions on the logit scale, losing power as the difference between them increases (Hauck & Donner 1977). We also assessed the performance of score tests (Morgan 2008, p. 102); however, some preliminary results (not reported here) indicated no benefit in using such tests.

As we briefly discussed in the introduction, reporting estimated CIs is more informative for inference than simply presenting the results of hypothesis tests. In line with this argument, some authors prefer to determine sample size based on the expected width of the CI for the measure of interest, which provides the analogue to power analysis in the significance testing framework: the larger the sample size, the narrower the CI and the larger the power of the test. It may seem that, with this approach, the concepts of power, null value and effect size to detect are avoided. However, the determination of sample size based on the precision of an estimation with no consideration of what constitutes a biologically significant difference is inappropriate (Greenland 1988; Daly 1991). Even when the focus is on estimation, whether or not biologically non-significant values are included in the CI is a central question for inference. When this is recognized, a correspondence is often made between the size of the (100-α)%-CI and the smallest difference that can be detected. It can be shown that this approach is effectively equivalent to setting the power of the test to 50% (Greenland 1988; Bristol 1989; Daly 1991; Goodman & Berlin 1994), a level that would often be considered too small. It can therefore be argued that the concept of power is still present, only that the decision on its level is implicit, while in power analysis, an explicit choice is made (Greenland 1988; Daly 1991). Note as well that the concept of power analysis can also be applied to the design of studies aimed at comparing occupancy estimated from a single sample to a chosen reference value.

While the ultimate aim of occupancy surveys is often to detect differences, very little has been published on associated design issues including sample size determination and trade-offs. Regarding multiple-season occupancy studies, two exceptions are MacKenzie (2005) and Field, Tyre & Possingham (2005), which explore scenarios involving a linear trend in occupancy on the logistic scale under the assumption of independence between seasons. MacKenzie (2005) presents a brief exploration showing how the coefficient of variation of the estimated trend parameter decreases as more seasons are added to the study (keeping the interval between seasons constant). Field, Tyre & Possingham (2005) use power analysis simulations to explore the optimal allocation of effort under a scenario involving three seasons and exponentially increasing survey costs. Our study (and formula) implies a scenario with two seasons. When the target number of seasons is moderately larger than two, sample size determination based on the change expected between the first and last season can still provide some useful (conservative) guidance, as we can normally expect the power not to be radically different to that obtained considering the trend across all seasons (e.g. Table 1, compare columns a, c). The frequency of surveying should be determined by the overall objectives of the study. In fact, if a linear trend is indeed expected and the focus is on estimating overall change, concentrating all the survey effort in the first and last seasons might provide a more powerful design (Table 1, columns a, e). Note nevertheless that a design with various sampling seasons may be more robust when there is noise because of variations from season to season on top of the trend and allows detecting departures from linearity.

Table 1. Power to detect a declining trend in occupancy (linear on the logit scale) for different designs ($K = 3$, $p = 0{\cdot}5$, $\alpha = 0{\cdot}05$). β_t is the rate of change on the logit scale, corresponding to a proportional decline R_4 between seasons 1 and 4.

Seasons			1, 2, 3, 4	1, 2	1, 4	1, 2	1, 4
Nr. of sites/season			S	S	S	$2S$	$2S$
R_4	ψ_1	β_t	(a)	(b)	(c)	(d)	(e)
0·5	0·8	−0·597	1·000	0·412	1·000	0·724	1·000
	0·5	−0·366	0·995	0·314	0·992	0·565	1·000
	0·2	−0·270	0·707	0·138	0·669	0·249	0·927
0·3	0·8	−0·382	0·981	0·158	0·970	0·327	1·000
	0·5	−0·206	0·748	0·136	0·705	0·223	0·945
	0·2	−0·143	0·294	0·074	0·275	0·101	0·491
0·15	0·8	−0·211	0·523	0·062	0·456	0·118	0·781
	0·5	−0·101	0·254	0·060	0·233	0·088	0·409
	0·2	−0·066	0·099	0·048	0·097	0·058	0·162

Different scenarios of initial occupancy ψ_1 and R_4 are assessed. Power is computed as the number of simulations out of 5000 in which the estimated $\hat{\beta}_t$ was significantly different from zero The trend is estimated based on (a) four seasons, (b, d) first and second season and (c, e) first and last season. In (a–c), $S = 200$ sites; in (d, e), all survey effort was concentrated in two seasons (i.e. 400 sites per season).

Our exploration illustrates how choosing an appropriate design in terms of the amount of replication vs. number of sampling sites optimizes the power of the study. We showed that choosing the amount of replication based on existing recommendations derived for single-season studies (MacKenzie & Royle 2005) will generally be also a good approach in terms of minimizing the effort required to achieve a given power. We assumed that all the individual surveys involve the same cost and assessed optimality based on total survey effort. However, in some scenarios, other cost functions might be more appropriate and the same exploration can be reproduced incorporating these. For instance, when assessing design trade-offs in terms of minimizing the asymptotic variance of the single-season occupancy estimator, MacKenzie & Royle (2005) consider a scenario in which the first visit to each site is more costly. Field, Tyre & Possingham (2005) incorporate a cost function that accounts for an exponentially increasing cost of adding new sites, to reflect a scenario in which sites with lowest access cost are chosen first (but note this sampling approach is not ideal as access cost and occupancy may not be independent).

It is important to realize that, while being a valuable study design tool, power analysis is an exploratory exercise rather than an exact science as it involves various assumptions and

decisions. This includes the fact that the system will often be more complex than described by the model, which implies that the true power may differ from the theoretical power. In practice, the design has to be determined based on assumed parameter values that are uncertain, and it is, therefore, always advisable to carry out sensitivity analysis to explore the impact of their variation. One needs also to decide on the effect sizes to consider. This should be evaluated case by case and comes back to a long-discussed issue: researchers should have a well-defined question before commencing a study and, therefore, an idea of what constitutes a biological effect of interest to them (Cherry 1998; Yoccoz, Nichols & Boulinier 2001). Finally, it is worth emphasizing that setting the significance level and target power should incorporate considerations about the seriousness of Type I vs. Type II error and, consequently, of their relative costs that can be biological, social or economic (Mapstone 1995; Di Stefano 2001, 2003; Field *et al*. 2004; Field, Tyre & Possingham 2005). While in some fields, the seriousness of Type I error is of greater concern, this is not a universal rule. Indeed, it might often be the contrary in ecological studies (Toft & Shea 1983; Shrader-Frechette & McCoy 1992). When dealing with threatened species, a small risk of missing a species decline will often be desired: the species going extinct is a high price to pay, as it is irreversible. Likewise, following again the precautionary principle, Type II error is of considerable concern when carrying out impact assessments. While estimating certain costs might be relatively straightforward (e.g. how much my monitoring programme costs?), quantifying others is unfortunately difficult (e.g. how much do we lose when a species goes extinct?). However, this should not justify disregarding the implications of the choice of α and β, to follow mechanically a default convention. An explicit choice should rather be made using the best knowledge available to ensure that the study is designed in the most meaningful way.

Acknowledgements

The authors were supported by an EPSRC/NCSE grant. JJL-M was also supported by the Centre for Ecology and Hydrology. The authors thank David Sewell for raising this problem, Martin Ridout and Byron Morgan for discussion and Matthew Spencer, Andrew Tyre and an anonymous reviewer for useful comments that helped improve the manuscript. We thank Jim Baldwin for pointing out a mis-statement in the earlier published version.

References

Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007) Sampling design trade-offs in occupancy studies with imperfect detection: examples and software. *Ecological Applications*, **17**, 281–290.

Bristol, D.R. (1989) Sample sizes for constructing confidence intervals and testing hypotheses. *Statistics in Medicine*, **8**, 803–811.

Casagrande, J.T., Pike, M.C. & Smith, P.G. (1978) An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, **34**, 483–486.

Cherry, S. (1998) Statistical tests in publications of The Wildlife Society. *Wildlife Society Bulletin*, **26**, 947–953.

Cochran, W.G. & Cox, G.M. (1957) *Experimental Designs*, 2nd edn. Wiley & Sons, Oxford.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences*, 2nd edn. Lawrence Erlbaum Associates, Hilldale, New Jersey, USA.

Cohen, J. (1990) Things i have learned (so far). *American Psychologist*, **45**, 1304–1312.

Daly, L. (1991) Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ*, **302**, 333–336.

Di Stefano, J. (2001) Power analysis and sustainable forest management. *Forest Ecology and Management*, **154**, 141–153.

Di Stefano, J. (2003) How much power is enough? Against the development of an arbitrary convention for statistical power calculations *Functional Ecology*, **17**, 707–709.

Di Stefano, J. (2004) A confidence interval approach to data analysis. *Forest Ecology and Management*, **187**, 173–183.

Dobson, A.J. & Gebski, V.J. (1986) Sample sizes for comparing two independent proportions using the continuity-corrected ArcSine transformation. *Journal of the Royal Statistical Society (series D)*, **35**, 51–53.

Donner, A. (1984) Approaches to sample size estimation in the design of clinical trials – a review. *Statistics in Medicine*, **3**, 199–214.

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

Field, S.A., Tyre, A.J., Jonzén, N., Rhodes, J.R. & Possingham, H.P. (2004) Minimizing the cost of environmental management decisions by optimizing statistical threholds. *Ecology Letters*, **7**, 669–675.

Fleiss, J.L. (1973) *Statistical Methods for Rates and Proportions*. Wiley & Sons, New York.

Fleiss, J.L., Tytun, A. & Ury, H.K. (1980) A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**, 343–346.

Goodman, S.N. & Berlin, J.A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200–206.

Gordon, I. & Watson, R. (1996) The myth of continuity-corrected sample size formulae. *Biometrics*, **52**, 71–76.

Greenland, S. (1988) On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology*, **128**, 231–237.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J.T. (2010) Design of occupancy studies with imperfect detection. *Methods in ecology and evolution*, **1**, 131–139.

Hauck, W.W. & Donner, A. (1977) Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, **72**, 851–853.

IUCN (2001) *Red List Categories and Criteria: Version 3.1*. IUCN Species Survival Commission, Gland.

Johnson, D.H. (1999) The Insignificance of Statistical Significance Test. *Journal of Wildlife Management*, **63**, 763–772.

Legg, C.J. & Nagy, L. (2006) Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management*, **78**, 194–199.

MacKenzie, D.I. (2005) What are the issues with presence absence data for wildlife managers? *Journal of Wildlife Management*, **69**, 849–860.

MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, New York.

Mapstone, B.D. (1995) Scalable decision rules for environmental impact studies: effect size, type I, and type II errors. *Ecological Applications*, **5**, 401–410.

McCarthy, M.A. & Possingham, H.P. (2007) Active adaptive management for conservation. *Conservation Biology*, **21**, 956–963.

Miller, D.A., Nichols, J.D., McClintock, B.T., Campbell Grant, E.H., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observation error occur: non-detection and species misidentification. *Ecology*, **92**, 1422–1428.

Morgan, B.J.T. (2008) *Applied Stochastic Modelling*, 2nd edn. Chapman & Hall, London.

Nickerson, R.S. (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, **5**, 241–301.

Shea, K. & Possingham, H.P. (2000) Optimal release strategies for biological control agents: an application of stochastic dynamic programming to population management. *Journal of Applied Ecology*, **37**, 77–86.

Shrader-Frechette, K.S. & McCoy, E.D. (1992) Statistics, costs and rationality in ecological inference. *Trends in Ecology & Evolution*, **7**, 96–99.
Steidl, R.J., Hayes, J.P. & Schauber, E. (1997) Statistical power analysis in wildlife research. *Journal of Wildlife Management*, **61**, 270–279.
Thomas, L. & Juanes, F. (1996) The importance of statistical power analysis: an example from Animal Behaviour. *Animal Behaviour*, **52**, 856–859.
Toft, C.A. & Shea, P.J. (1983) Detecting community-wide patterns: estimating power strengthens statistical inference. *The American Naturalist*, **122**, 618–625.
Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.
Ury, H.K. & Fleiss, J.L. (1980) On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics*, **36**, 347–351.
Vaeth, M. (1985) On the use of Wald's test in exponential families. *International statistical review*, **53**, 199–214.
Vorburger, M. & Munoz, B. (2006) Simple power calculations: how do we know we are doing them the right way? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 3809–3812. American Statistical Association, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/.
Walters, D.E. (1979) In defence of the ArcSine approximation. *Journal of the Royal Statistical Society (series D)*, **28**, 219–222.
Yoccoz, N.G. (1991) Use, overuse, and misuse of significance test in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72**, 106–111.
Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, **16**, 446–453.

Received 3 March 2012; accepted 30 April 2012
Handling Editor: Robert Freckleton

Supporting information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Supporting figures and example.

Appendix S2. Two-season occupancy model with Markovian dependence. [Correction added on 14 Sept after first online publication: Replacement file supplied for Appendix S2.]

Appendix S3. R code for power analysis.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.