Ecology, 83(9), 2002, pp. 2387–2393

HOW SHOULD DETECTION PROBABILITY BE INCORPORATED INTO ESTIMATES OF RELATIVE ABUNDANCE?

DARRYL I. MACKENZIE[1,3] AND WILLIAM L. KENDALL[2]

[1]*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 USA*
[2]*USGS Patuxent Wildlife Research Center, 11510 American Holly Drive, Laurel, Maryland 20708-4017 USA*

Abstract. Determination of the relative abundance of two populations, separated by time or space, is of interest in many ecological situations. We focus on two estimators of relative abundance, which assume that the probability that an individual is detected at least once in the survey is either equal or unequal for the two populations. We present three methods for incorporating the collected information into our inference. The first method, proposed previously, is a traditional hypothesis test for evidence that detection probabilities are unequal. However, we feel that, a priori, it is more likely that detection probabilities are actually different; hence, the burden of proof should be shifted, requiring evidence that detection probabilities are practically equivalent. The second method we present, equivalence testing, is one approach to doing so. Third, we suggest that model averaging could be used by combining the two estimators according to derived model weights. These differing approaches are applied to a mark–recapture experiment on Nuttall's cottontail rabbit (*Sylvilagus nuttallii*) conducted in central Oregon during 1974 and 1975, which has been previously analyzed by other authors.

Key words: bioequivalence; capture–recapture; detection probability; equivalence testing; growth rate; hypothesis testing; model averaging; Nuttall's cottontail rabbit; relative abundance; Sylvilagus nuttallii.

INTRODUCTION

The change in population size over time or space is of interest to ecologists in several settings. Much of matrix modeling in population ecology is devoted to determining the growth rate of a population under various conditions (see Caswell 2001). In conservation biology, much attention is devoted to estimating trends, in which a decline would indicate a population that might require human intervention to avoid entering a threatened or endangered status (Brindley et al. 1998, Sugimura et al. 2000). In some cases of harvest management, regulations are based largely on estimates of population change over time (e.g., Dolton et al. 2001, Kelley 2001).

Interest in change over space or time is not limited to populations. Its counterpart is found in community ecology, where number of species rather than abundance of animals is the issue, and the focus is on the change in species richness over time or space (Nichols et al. 1998*a*, *b*). Our discussion of relative abundance in this paper is directly applicable to this case.

Whether defined over space (relative population size) or time (population growth rate), relative abundance between two points (λ) is defined as

$$\lambda = \frac{N_2}{N_1} \tag{1}$$

where N_1 and N_2 are the sizes of the populations at points 1 and 2, respectively.

Of course population size is rarely, if ever, known. Often a survey is conducted that yields a count of animals at each point (C_i). These counts may serve as an index to abundance, and together yield an estimate of relative abundance:

$$\hat{\lambda}_1 = \frac{C_2}{C_1}. \tag{2}$$

However, this estimator will be biased if detection probabilities are different at points 1 and 2. Detection probability, P_i, is the probability that a member of the population of interest is included in the count at time or location *i*. The expected value of the count is

$$E[C_i] = N_i \times P_i. \tag{3}$$

If detection probability can be estimated, i.e., using capture–recapture methods (see Otis et al. 1978, Hug-

Manuscript received 7 December 2001; revised 31 January 2002; accepted 3 February 2002.

[3] Present address: Proteus Research and Consulting Ltd., P.O. Box 5193, Dunedin, New Zealand.
E-mail: darryl@proteus.co.nz

gins 1991) or distance sampling (Buckland et al. 1993), a more robust estimator, derived from Eq. 3, is

$$\hat{\lambda}_2 = \frac{C_2/\hat{P}_2}{C_1/\hat{P}_1}. \qquad (4)$$

We agree wholeheartedly with previous authors (e.g., Skalski et al. 1983, Thompson et al. 1998, Nichols et al. 2000, Anderson 2001, Yoccuz et al. 2001) that estimating relative abundance based on indices alone (e.g., raw counts) is naïve, and that every effort should be made to design a survey that enables detection probabilities to be estimated. A pertinent question is how the estimation of detection probability, an important parameter statistically, should be incorporated into the estimation of relative abundance, the parameter of ecological interest. That is the focus of this paper. We begin with an approach described in Skalski et al. (1983) that involves a hypothesis test to determine whether the P's can be excluded from the estimation of λ. We offer two alternatives to their approach. The first is a hypothesis-testing approach based on equivalence tests (e.g., see Manly 2001). The second is a model-averaging approach, as in Buckland et al. (1997) and Burnham and Anderson (1998), which argues that relative abundance should be estimated as a weighted average of $\hat{\lambda}_1$ and $\hat{\lambda}_2$. The relative weights for each model could be determined via information-theoretic methods, such as Akaike's Information Criterion (AIC), or bootstrapping (Burnham and Anderson 1998). We advocate the model-averaging approach for estimation, but argue that equivalence testing is the most logical approach to consider when there is direct interest in determining whether detection probabilities for the two populations are practically equivalent.

Incorporating Detection Probabilities

Traditional hypothesis-testing approach

Skalski et al. (1983; also see Skalski and Robson 1992:97) emphasized the importance of estimating P's, but advocated that they should be ignored if they could not be shown to be different, as $\hat{\lambda}_1$ would then be an appropriate estimator with smaller standard error. Skalski et al. (1983) outlined two approaches for evaluating the following hypotheses:

$$H_0\colon \quad \mathbf{p}_1 = \mathbf{p}_2$$

$$H_A\colon \quad \mathbf{p}_1 \neq \mathbf{p}_2$$

where $\mathbf{p}_i$ is a vector of capture probabilities obtained from a capture–recapture experiment at point *i*. One method was a likelihood ratio test and the other was a contingency table approach. Under either approach, if the null hypothesis, H_0, is rejected, then $\hat{\lambda}_2$ should be used to estimate relative abundance; otherwise, $\hat{\lambda}_1$ should be used. Nichols et al. (1998*a, b*) applied this approach in estimating change in species richness over space and time, and have incorporated both options into program COMDYN (Hines et al. 1999).

However, the burden of proof is on H_A (detection probabilities are assumed equal until sufficient evidence suggests otherwise), yet in most practical situations, it is often unreasonable to expect detection probabilities to be constant. For instance, detection probabilities may vary with (1) environmental variables such as weather conditions; (2) different observers; or (3) local habitats. Hence, a priori, a null hypothesis of no difference is likely to be false. The testing of trivial null hypotheses in ecological settings has recently come under a great deal of criticism (Steidl et al. 1997, Johnson 1999, Anderson et al. 2000), and in such situations, low power of the testing procedure is most probably the cause of H_0 not being rejected. Conversely, with very large sample sizes, H_0 may be rejected even if a practically inconsequential difference exists. Failing to reject a null hypothesis of ''no difference'' indicates that there is insufficient evidence that detection probabilities are different, but does not prove that they are the same.

Skalski et al. (1983) acknowledged the potential lack of robustness of $\hat{\lambda}_1$ and also that the low power of the testing procedure may result in $\hat{\lambda}_1$ being used inappropriately. They suggested raising the α level of the test to 0.10, to improve the power of the test. This is a reasonable approach to minimize Type II errors, but an alternative is to shift the burden of proof to find evidence of homogeneity, using equivalence testing.

Equivalence-testing approach

Equivalence tests may be constructed in one of two ways, reflecting the hypothesis that is to be tested. One may use an equivalence hypothesis, which assumes that two (or conceptually more) model parameters are equivalent to within some pre-specified bounds; the testing procedure seeks evidence to falsify this. Traditional hypothesis tests could be considered as a special case of such an approach, where the pre-specified bounds are so small that they amount to exact equality. Alternatively, one may assume a hypothesis of inequality and seek evidence that the parameters of interest are practically equivalent. It is this second form of equivalence tests that we shall focus on, as this represents our a priori beliefs that detection probabilities should be assumed to differ by a substantial amount until sufficient evidence is gathered to the contrary. This shifting of the burden of proof is appropriate here, because $\hat{\lambda}_1$ could be badly biased when detection probabilities are unequal. The value of such tests has long been recognized in pharmaceutical and medical situations. Early examples in these fields include Westlake (1973), Metzler (1974), and Dunnett and Gent (1977). More recently, the applicability of equivalence tests to environmental problems has been recognized (McBride et al. 1993, Erickson and McDonald 1995, McBride 1999, Manly 2001), but to date, there is little evidence that the methods, and mind set, have filtered into ecological studies.

With this class of test, the hypothesis to be tested assumes that the items of interest are substantially different; the alternative is that they are equivalent (i.e., the true difference lies within some pre-specified bounds). For example, the hypothesis (H) and its alternative (K) for testing whether the parameter μ is inequivalent for two populations would be:

$$H\colon \quad \mu_1 - \mu_2 \le \theta_L \qquad \mu_1 - \mu_2 \ge \theta_U$$

$$K\colon \quad \theta_L < \mu_1 - \mu_2 < \theta_U$$

where θ_L and θ_U are the accepted lower and upper limits for the true difference, between which the parameters will be deemed equivalent. This hypothesis can be tested using two one-sided tests (e.g., see Berger and Hsu 1996, Manly 2001):

$$H_1\colon \quad \mu_1 - \mu_2 \le \theta_L \qquad H_2\colon \quad \mu_1 - \mu_2 \ge \theta_U$$

$$K_1\colon \quad \mu_1 - \mu_2 > \theta_L \qquad K_2\colon \quad \mu_1 - \mu_2 < \theta_U.$$

The overall hypothesis is rejected at the $100\alpha\%$ level if and only if both of the subhypotheses (H_i) for the one-sided tests are rejected, each at the $100\alpha\%$ level. No adjustment for multiple comparisons is required when testing for equivalence with this type of method (Berger and Hsu 1996). The overall P value for the equivalence test is the maximum of the P values obtained from the tested subhypotheses (Berger and Hsu 1996).

Manly (2001:184–190) presents the equivalence test using two one-sided t tests. However, for non-normal data, such as counts and detection probabilities, one-sided likelihood ratio tests may be more appropriate. A one-sided likelihood ratio test for H_1 could be conducted by: (1) obtaining the log-likelihood for the assumed model (l_H) where $\mu_1 = \mu_2 + \theta_L$; (2) obtaining the log-likelihood for the alternative model (l_K) constraining $\mu_1 > \mu_2 + \theta_L$; (3) calculating the test statistic: $X^2 = 2(l_K - l_H)$; and (4) comparing X^2 to the $\bar{\chi}^2_1$ distribution to determine whether X^2 is unusually large.

Asymptotically, X^2 will have a chi-bar-square distribution with 1 degree of freedom, $\bar{\chi}^2_1$, which is a 50:50 mixture of zeros and χ^2_1 random variates (Dykstra and El Barmi 1997). More generally, the degrees of freedom associated with X^2 and $\bar{\chi}^2$ will be the difference in the number of parameters estimated in the tested and alternative models. This general procedure would be repeated for the second one-sided test, H_2, and some overall conclusion about the equivalence of the parameters could be made.

Key to the concept of equivalence testing is: for what range of values should two parameters be deemed equivalent? Clearly, the appropriate bounds will vary with each application, but they should be chosen to have some reasonable biological or management interpretation. In the case of relative abundance, one might be willing to tolerate a 10% bias in λ due to using $\hat{\lambda}_1$ for estimation. In a simple case where $P_i = p_i$, from Eq. 4 we can see that such a bias would result if $p_2/p_1 < 0.90$ or $p_2/p_1 > 1.10$. Some idea about the value of p_1 or p_2 would then be needed. If $p_1 \approx 0.5$, then this tolerance would be exceeded if $p_2 < 0.45$ or $p_2 > 0.55$. In this case, therefore, $\theta_L = -0.05$ and $\theta_U = 0.05$. This exercise of determining the value of θ is similar to the determination of a meaningful effect size required for calculations of power when designing a study. Although some may hesitate to use equivalence tests because of the subjective nature of defining the appropriate bounds, we strongly argue that this is preferable to the testing of the ''objective'' hypothesis of exact equality, which is very unlikely to be true from the outset.

Model-averaging approach

The choice between traditional tests of homogeneity of detection probabilities and equivalence tests could depend on philosophy, or a relative emphasis on bias vs. precision based on a metric like MSE (mean square error). An alternative approach to balancing bias and precision is to forego testing altogether and use model averaging to estimate relative abundance. Much has been written in recent years about selecting among statistical models using measures like AIC and model averaging (e.g., see Norris and Pollock 1996, Buckland et al. 1997, Burnham and Anderson 1998). Under this approach, a model that includes equal P's can be considered for the sake of parsimony (i.e., balancing the potential bias due to that constraint against the increased precision expected by reducing the number of parameters considered), along with more general models. The relative weight of evidence for each model can be obtained from either the difference in AIC (or AIC_C, which includes a correction for small sample bias) values for the fitted models or via bootstrapping.

Based upon AIC values, the weight for the jth of the m models fitted to the data, w_j, is

$$w_j = \frac{\exp(-\Delta\text{AIC}_j)}{\sum_{k=1}^{m} \exp(-\Delta\text{AIC}_k)} \tag{5}$$

where ΔAIC_j is the difference in AIC between the minimum value and the value for model j (Burnham and Anderson 1998). Alternatively, using a bootstrapping approach, w_j would be determined by the proportion of times that model j was the most parsimonious, when the m models are fit to each of the B bootstrap resampled data sets (Norris and Pollock 1996).

It follows that model-averaged estimates of relative abundance and associated standard error would be as follows (Burnham and Anderson 1998):

$$\bar{\lambda} = \sum_{j=1}^{m} w_j \hat{\lambda}_j \tag{6}$$

$$\text{SE}(\bar{\lambda}) = \sum_{j=1}^{m} w_j \sqrt{\text{SE}(\hat{\lambda}_j)^2 + (\hat{\lambda}_j - \bar{\lambda})^2}. \tag{7}$$

TABLE 1. Collected data on Nuttall's cottontail rabbit (*Sylvilagus nuttallii*).

Capture history	1974	1975
11	7	17
10	80	84
01	7	10
Total no. unique animals	94	111

Note: Capture histories represent the possible observable combinations of rabbits encountered (1) and not encountered (0) at each trapping occasion.

Such an estimation approach is appealing. Whereas testing of biological hypotheses, especially based on planned experiments, is logical and is part of scientific methodology, inference about nuisance parameters such as detection probability is merely a step toward scientific inference, and therefore testing here is less appealing. If testing for the homogeneity of the P's concludes that they are not equivalent, then $\hat{\lambda}_2$ is used and relative abundance is estimated in the face of uncertainty about the value of the P's or, more importantly, their ratio. However, if testing concludes that the P's are equivalent, then the ratio of the P's is assumed to be 1.0 and the uncertainty in that value is ignored. Therefore, the ecological inference of interest is predicated on the validity of a previous statistical hypothesis test (the exact equality of the P's), without incorporating the error rate of that test into the error rate of the ecological inference. Even Skalski et al. (1983) advocate hypothesis testing to determine whether or not to consider P's in the calculation of λ, but advocate using estimation instead of testing to make the ecological inference about λ.

Taking a model-averaging approach retains the uncertainty in λ due to model selection, and balances the trade-off between bias and precision in using $\hat{\lambda}_1$ or $\hat{\lambda}_2$. The precision that results from this process incorporates the uncertainty in the value of the ratio of the P's.

EXAMPLE: NUTTALL'S COTTONTAIL RABBIT

We demonstrate the use of equivalence tests and model averaging by reassessing the Nuttall's cottontail rabbit (*Sylvilagus nuttallii*) data used by Skalski et al. (1983). In August of 1974 and 1975, the population on an 87.0-ha study area in central Oregon was sampled using capture–recapture methods. Cottontails were live-trapped, marked with picric acid, and released. On 5 September 1974 and 3 September 1975, cottontails were ''recaptured'' during drive censuses, with the resulting data summarized in Table 1. Using this capture–recapture design, the probability of being detected at least once is

$$P_i = 1 - (1 - p_{1,i})(1 - p_{2,i}) \tag{8}$$

where $p_{t,i}$ is the capture probability at sampling occasion t in year i. Skalski et al. (1983) acknowledge that it was possible that the same animal may have been sighted multiple times during the drive censuses, but they use the data for illustrative purposes only, as we do here.

To make inference about the p's, we used Huggins' (1991) model for closed populations that conditions the likelihood upon the animals detected at least once. This does not enable population size to be estimated directly, but here we wished to focus on the p's. Similar conclusions are likely to be made if other appropriate closed population models were used (e.g., see Otis et al. 1978), but Huggins (1991) method was used for its practical convenience. Table 2 contains the estimated p's for the two models considered.

Based upon these values, our estimates of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are

$$\hat{\lambda}_1 = \frac{111}{94} = 1.18$$

$$\hat{\lambda}_2 = \frac{111/[1 - (1 - 0.63)(1 - 0.17)]}{94/[1 - (1 - 0.50)(1 - 0.08)]} = 0.92$$

which, allowing for round off error, agree with those of Skalski et al. (1983) who gave estimates of relative abundance in the two years as 1.18 ± 0.10 and 0.91 ± 0.27; (estimate ± 1 SE). Clearly, by accounting for detection probabilities, the estimate of relative abundance is less precise, but note that the implications of the point estimates themselves differ substantially, indicating an increasing and decreasing population, respectively.

Taking a traditional hypothesis-testing approach, Skalski et al. (1983) concluded that there was insufficient evidence to reject the null hypothesis that the p's were equal ($\chi^2 = 3.4$, df = 2, P value = 0.18), and hence suggested that $\hat{\lambda}_1$ was an appropriate estimator.

However, a different conclusion is reached if equivalence testing is used to determine whether there is sufficient evidence to reject the hypothesis that the p's are inequivalent. The intent, therefore, is to test the following hypothesis, where $\mathbf{p}_{74}$ and $\mathbf{p}_{75}$ are the vectors of capture probabilities for the population in each year, with θ_L and θ_U being the vectors that represent the bounds on $\mathbf{p}_{74} - \mathbf{p}_{75}$:

$$H\colon\quad \mathbf{p}_{74} - \mathbf{p}_{75} \le \boldsymbol{\theta}_L \qquad \mathbf{p}_{74} - \mathbf{p}_{75} \ge \boldsymbol{\theta}_U$$

$$K\colon\quad \boldsymbol{\theta}_L < \mathbf{p}_{74} - \mathbf{p}_{75} < \boldsymbol{\theta}_U.$$

Because there were two trapping occasions in each

TABLE 2. Estimated mean capture probabilities (and 1 SE) for each of the two models considered for the cottontail rabbits example using Huggins' (1991) conditional likelihood model.

Trapping occasion	Capture probabilities		
		Unequal	
	Equal $\mathbf{p}_.$	$\mathbf{p}_{74}$	$\mathbf{p}_{75}$
1	0.59 (0.08)	0.50 (0.13)	0.63 (0.09)
2	0.13 (0.02)	0.08 (0.03)	0.17 (0.04)

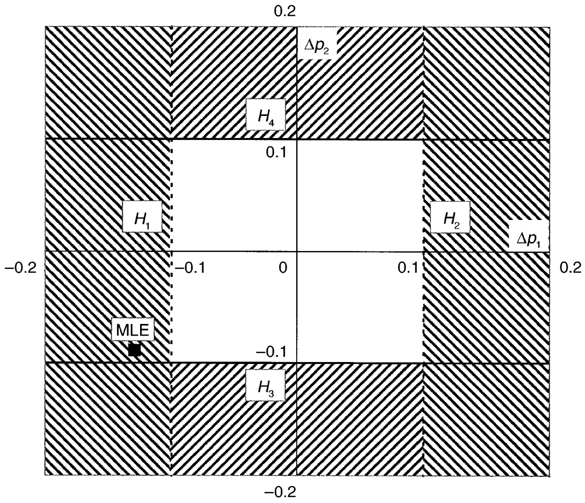

FIG. 1. Illustration of the four subhypotheses for the equivalence test in the Nuttall's cottontail rabbit example. Shaded (hatched) areas represent the regions where the hypotheses (H_j) are true; the unshaded central area is the intersection of four rejection regions: the equivalence region. Axes are the difference between years (1974 and 1975) of the capture probabilities for the first and second sampling occasion ($\Delta p_t = p_{t,74} - p_{t,75}$). The difference between the maximum likelihood estimates (MLE) for the example is also indicated.

year, to reject the overall hypothesis that the *p*'s are inequivalent, each of the four following subhypotheses must be rejected:

$$H_1\colon \quad p_{1,74} \leq p_{1,75} + \theta_{1,\mathrm{L}}$$

$$H_2\colon \quad p_{1,74} \geq p_{1,75} + \theta_{1,\mathrm{U}}$$

$$H_3\colon \quad p_{2,74} \leq p_{2,75} + \theta_{2,\mathrm{L}}$$

$$H_4\colon \quad p_{2,74} \geq p_{2,75} + \theta_{2,\mathrm{U}}.$$

We used $\theta_{t,\mathrm{L}} = -0.1$ and $\theta_{t,\mathrm{L}} = 0.1$. These limits for determining the practical equivalence of the *p*'s are very liberal, because a true absolute difference of 0.1 would create a bias of approximately ±20% if $\hat{\lambda}_1$ were used to estimate relative abundance. Fig. 1 illustrates how the intersection of the rejection regions for these four hypotheses creates the equivalence region: the area of the parameter space where the *p*'s will be deemed equivalent.

Table 3 presents the results of the likelihood ratio tests. Because the *P* value for the overall test is 0.50, there is no evidence to reject the hypothesis that the *p*'s are inequivalent; hence, one would conclude that $\hat{\lambda}_2$ is the appropriate estimator to use. Alternatively, because the difference in the maximum likelihood estimates lies within the shaded region of H_1 (Fig. 1), there will not be sufficient evidence to reject H_1. Hence, for this example, the same conclusion could be reached without performing any actual tests.

Avoiding hypothesis testing altogether and taking a model-averaging approach provides some further interesting insight. Using Program MARK (White and Burnham 1999), the $\mathrm{AIC_c}$ model weights are 0.57 and 0.43 for models with equal and unequal *p*'s, respectively. This suggests that although the majority of support is for the model with equal *p*'s, it is by no means overwhelming and there is also a large degree of support for the unequal detection probability model. The lack of a clear choice for a single model reflects the contrasting results obtained from the two hypothesis-testing approaches. From Eqs. 6 and 7, the averaged estimate of relative abundance is $\bar{\lambda} = 1.06$ (SE = 0.22). Therefore, using model averaging, we obtain an estimate of relative abundance that is between $\hat{\lambda}_1$ and $\hat{\lambda}_2$; hence, it is probably less biased than $\hat{\lambda}_1$ and has a smaller standard error than $\hat{\lambda}_2$. We would conclude that the relative abundance of cottontails in 1975 is similar to that in 1974.

DISCUSSION

The relative merits of the three methods detailed here depend upon the overall goal of the study and a priori beliefs of the researchers. If there is scientific interest in the *P*'s, and a decision between models with equal or unequal detection probabilities is required and appropriate, hypothesis testing may be more appropriate than a model-averaging approach. Given the recent arguments against the testing of the trivial hypothesis of exact equality used by traditional hypothesis tests, one could suggest that such tests should be abandoned altogether in favor of equivalence tests. Consideration of where the burden of proof should lie then dictates which hypothesis (equivalence or inequivalence) should be tested. We believe that, in settings similar to those we have just described, equivalence tests using the inequivalence hypothesis are the most appropriate method to use. Falsely concluding that the *P*'s are about equal, by failing to reject a hypothesis of equivalence, will lead to biased estimates of relative abundance.

However, we advocate an estimation approach to inference on *P*'s and, within that approach, we encourage model averaging because it incorporates not only the level of uncertainty in the estimate for a given model, but also the uncertainty in the model selection itself. We have illustrated model averaging with a reasonably simple case. For more complex problems, one could

TABLE 3. Results for determining whether capture probabilities at each sampling occasion are equivalent, to within ±0.1, in 1974 and 1975.

Subhypothesis	l_{H}	l_{K}	X^2	*P* value
H_1: $p_{1,74} \leq p_{1,75} - 0.1$	−128.66	−128.66	0.00	0.50
H_2: $p_{1,74} \geq p_{1,75} + 0.1$	−129.65	−128.64	2.02	0.08
H_3: $p_{2,74} \leq p_{2,75} - 0.1$	−128.68	−128.64	0.07	0.40
H_4: $p_{2,74} \geq p_{2,75} + 0.1$	−135.60	−128.64	13.91	<0.01

Note: l_{H} and l_{K} are the log-likelihood values under the assumed and alternative models; X^2 is the test statistic; and the *P* value is determined from the chi-bar-square distribution with 1 degree of freedom.

easily envisage a suite of potentially realistic models that could be fit to the data, each with different estimates of relative abundance. In such a case, model averaging would seem to be a reasonable method of estimating relative abundance in the face of model selection uncertainties.

We stress that, beyond the information required to estimate detection probabilities, neither equivalence testing or model averaging involves the collection of additional data from the field. They are merely different methods of analysis that can be conducted, once the survey is completed, in order to make appropriate inference about relative abundance.

Strictly speaking, in many situations such as the cottontail example, performing hypothesis tests upon the capture probability vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, either using traditional or equivalence approaches, does not address the main question of interest: is the overall probability of being included in the count, P_i, about the same at both points? Often P_i will be a nonlinear function of the elements in $\mathbf{p}_i$, so in some circumstances the P_i's may be equal even when the vectors $\mathbf{p}_i$ are not. For instance, in a two-sample capture–recapture experiment where $\mathbf{p}_1 = (0.7, 0.3)$ and $\mathbf{p}_2 = (0.3, 0.7)$, although the vectors are different, P_1 and P_2 both equal 0.79; hence, $\hat{\lambda}_1$ would be an appropriate estimator. However, we believe that such situations are likely to be rare, and would be clearly evident when one considers the parameter estimates. In this example, the estimated p's do not suggest this to be the case. Potential solutions to this problem are an ongoing area of research.

Our use of equivalence tests here has focused on inference of detection probabilities. However, the availability of this tool adds another dimension to considering inferences of direct scientific interest in a hypothesis-testing context. In comparing two populations with respect to some vital rate (e.g., survival rate), there is now an option for which hypothesis should bear the burden of proof, and the question of equality can be addressed not in the trivial terms of exact equality, but as being within some meaningful difference θ. There are precedents for this in other related fields. The switching of the burden of proof has long been used in the pharmaceutical industry, in which companies must prove that new drugs are equivalent to existing formulations. Also, following the cleanup of a contaminated site, the U.S. Environmental Protection Agency requires the responsible parties to prove that the cleaned site is equivalent to an undisturbed one (Manly 2001). In ecology, Sauer and Link (2002) take a Bayesian estimation approach to defining equality within a given tolerance. They define stability as a population growth rate within a specified distance θ from 1.0, and compute the proportion of species in the North American Breeding Bird Survey whose populations are stable. A similar approach could be taken with equivalence tests.

Our premise has been that detection probability tends to vary over time and space, and therefore the burden of proof should be on showing that it is equivalent, not the opposite. Experience has confirmed differences in detection probability more often than not; therefore, the default approach in designing any monitoring program should be to estimate detection probability and adjust estimates of relative abundance accordingly. If equivalence testing or model selection indicates that the differences in detection probabilities are trivial, this can be incorporated into the analysis as previously described. Nevertheless, the practitioner should resist the temptation to use a result of equivalence from a single analysis to conclude that future estimation of detection probabilities is unnecessary. Just as populations and communities are dynamic, so are the factors that determine the values of detection probabilities (e.g., observers, weather, animal activity).

The methods required to estimate detection probability are often seen to be expensive or difficult to implement in the field. Of course in many cases, animals must be captured and marked, which can be difficult and expensive, but feasible. In other cases, there is no need to physically capture animals (e.g., Karanth and Nichols 1998, Langtimm et al. 1998, Nichols et al. 2000, Boyce et al. 2001, MacKenzie et al. 2002). Another argument has been that simple index methods (e.g., raw counts) require fewer or none of the assumptions required by complex capture–recapture methods. We have shown that the opposite is true: $\hat{\lambda}_1$ actually requires more restrictive assumptions than $\hat{\lambda}_2$ that are unlikely to be true in practice. Whether detection probability is easy or difficult to estimate, without doing so, it is impossible to determine whether a change in the number of individuals counted by surveys conducted at various points in time or space is due to a change in the size of the populations or to changes in detection probabilities.

Acknowledgments

We thank J. Nichols, J. Sauer, and D. Fletcher for their helpful suggestions on an earlier draft of this paper, and S. Buckland and G. McBride for their excellent comments.

Literature Cited

Anderson, D. R. 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin **29**:1294–1297.

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

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

Boyce, M. S., D. I. MacKenzie, B. F. J. Manly, M. A. Haroldson, and D. Moody. 2001. Negative binomial models for abundance estimation of multiple closed populations. Journal of Wildlife Management **65**:498–509.

Brindley, E., K. Norris, T. Cook, S. Babbs, C. F. Brown, P. Massey, R. Thompson, and R. Yaxley. 1998. The abundance and conservation status of redshank *Tringa totanus* nesting on saltmarshes in Great Britian. Biological Conservation **86**:289–297.

Buckland, S. T., D. R. Anderson, K. P. Burnham, and J. L.

Laake. 1993. Distance sampling: estimating abundance of biological populations. Chapman and Hall, London, UK.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: an integral part of inference. Biometrics **53**:603–618.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference. Springer-Verlag, New York, New York, USA.

Caswell, H. 2001. Matrix population models. Second edition. Sinauer Associates, Sunderland, Massachusetts, USA.

Dolton, D. D., R. D. Holmes, and G. W. Smith. 2001. Mourning dove breeding population status, 2001. U.S. Fish and Wildlife Service, Laurel, Maryland, USA.

Dunnett, C. W., and M. Gent. 1977. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. Biometrics **33**: 593–602.

Dykstra, R., and H. El Barmi. 1997. Chi-bar-square distributions. Pages 89–93 *in* S. Kotz, C. B. Read, and D. L. Banks, editors. Encyclopedia of statistical sciences. Update volume 1. John Wiley, New York, New York, USA.

Erickson, W. P., and L. L. McDonald. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. Environmental Toxicology and Chemistry **14**: 1247–1256.

Hines, J. E., T. Boulinier, J. D. Nichols, J. R. Sauer, and K. H. Pollock. 1999. COMDYN: software to study the dynamics of animal communities using a capture–recapture approach. Bird Study **46**(supplement):209–217.

Huggins, R. M. 1991. Some practical aspects of a conditional likelihood approach to capture experiments. Biometrics **47**: 725–732.

Johnson, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management **63**:763–772.

Karanth, K. U., and J. D. Nichols. 1998. Estimation of tiger densities in India using photographic captures and recaptures. Ecology **8**:2852–2862.

Kelley, J. R., Jr. 2001. American woodcock population status, 2001. U.S. Fish and Wildlife Service, Laurel, Maryland, USA.

Langtimm, C. A., T. J. O'Shea, R. Pradel, and C. A. Beck. 1998. Estimates of annual survival probabilities for adult Florida manatees (*Trichechus manatus latirostrus*). Ecology **79**:981–997.

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

Manly, B. F. J. 2001. Statistics for environmental science and management. Chapman and Hall/CRC, Boca Raton, Florida, USA.

McBride, G. B. 1999. Equivalence tests can enhance environmental science and management. Australian and New Zealand Journal of Statistics **41**:19–29.

McBride, G. B., J. C. Loftis, and N. C. Adkins. 1993. What do significance tests really tell us about the environment? Environmental Management **17**:423–432.

Metzler, C. M. 1974. Bioavailability—a problem of equivalence. Biometrics **30**:309–317.

Nichols, J. D., T. Boulinier, J. E. Hines, K. H. Pollock, and J. R. Sauer. 1998*a*. Inference methods for spatial variation in species richness and community composition when not all species are detected. Conservation Biology **12**:1390–1398.

Nichols, J. D., T. Boulinier, J. E. Hines, K. H. Pollock, and J. R. Sauer. 1998*b*. Estimating rates of local species extinction, colonization, and turnover in animal communities. Ecological Applications **8**:1213–1225.

Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon, and P. J. Heglund. 2000. A double-observer approach for estimating detection probability and abundance from counts. Auk **117**:393–408.

Norris, J. L., and K. H. Pollock. 1996. Including model uncertainty in estimating variance in multiple capture studies. Environmental and Ecological Statistics **3**:235–244.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs **62**.

Sauer, J. R., and W. A. Link. 2002. Hierarchical modeling of population stability and species group attributes from survey data. Ecology **83**:1743–1751.

Skalski, J. R., and D. S. Robson. 1992. Techniques for wildlife investigations. Academic Press, San Diego, California, USA.

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

Steidl, R. J., J. P. Hayes, and E. Schauber. 1997. Statistical power analysis in wildlife research. Journal of Wildlife Management **61**:270–279.

Sugimura, K., S. Sato, F. Yamada, S. Abe, H. Hirakawa, and Y. Handa. 2000. Distribution and abundance of the Amami rabbit *Pentalagus furnessi* in the Amami and Tokuno Islands, Japan. Oryx **34**:198–206.

Thompson, W. L., G. C. White, and C. Gowan. 1998. Monitoring vertebrate populations. Academic Press, San Diego, California, USA.

Westlake, W. J. 1973. Use of statistical methods in evaluation of in vivo performance of dosage forms. Journal of Pharmaceutical Sciences **62**:1579–1589.

White, G. C., and K. P. Burnham. 1999. Program MARK for survival estimation. Bird Study **46**(supplement):120–139.

Yoccuz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. Trends in Ecology and Evolution **16**:446–453.