Studies in Avian Biology No. 29:168–172

SOME CONSEQUENCES OF USING COUNTS OF BIRDS BANDED AS INDICES TO POPULATIONS

JOHN R. SAUER AND WILLIAM A. LINK

Abstract. In mist-net studies, it is often difficult to use capture–recapture methods to estimate number of birds present. Many investigators use number of birds captured as an index of population size. We investigate the consequences of using indices of bird abundance as surrogates for population size in hypothesis tests. Unless all of the birds present are captured, indices are biased estimates of local population size, and the amount of bias depends on the proportion of birds captured. We demonstrate the potential effects of bias on hypothesis tests based on indices. The bias generally causes type I error rates to be inflated. Investigators should either estimate the proportion of animals captured using capture–recapture methods or demonstrate that results of hypothesis tests based on indices are not consequences of bias in the indices.

Key Words: abundance estimation, banding, bias, capture–recapture, counts, index, population size

Banding data provide the only source of information regarding many interesting questions about bird populations. Data from mist-net studies are presently used to estimate population trends of passerine birds (Dawson 1990, Hussell et al. 1992), to examine survival and population sizes of birds (e.g., Faaborg and Arendt 1992b), and to evaluate productivity of passerines (DeSante 1992). Large-scale banding programs such as MAPS (DeSante 1992) and the British Constant Effort Sites (Peach 1993) provide the opportunity for monitoring trends and demographic characteristics at regional geographic scales.

Unfortunately, in mist-net studies, relatively few individuals of the target species are typically encountered. Because mist nets have a limited height, the probability of capturing a bird that does not forage in the understory is relatively small. Also, after being captured, birds may become aware of the location of nets, leading to low recapture rates (DeSante 1992). Consequently, most bird species are represented by small sample sizes from any study site.

Small sample sizes pose many challenges for analysts of mist-net data. The most important problem relates to use of capture-recapture methods with small samples. These methods provide many interesting opportunities for estimation of demographic parameters (Kendall et al. *this volume*), but small samples can preclude estimation from individual sites or greatly lower the power of tests for differences in parameters over time or between sites. Many investigators choose to avoid the problems inherent in small-sample capture–recapture analyses by using indices in their population analyses. For example, the total number of birds captured at a site is used as an index to total population size, trends are estimated based on changes in the total capture indices, total numbers of recaptures are used as an index of return (or survival) rates, and the ratio of number of young to adults captured is used as an index of productivity.

In this paper, we explore the consequences of using indices in analysis. We develop a conceptual framework for analyzing indices and relating them to possible changes in the underlying populations. Finally, we demonstrate how indices should be considered in terms of underlying capture–recapture models.

WHAT IS AN INDEX?

An index count is often defined as any kind of count that reflects the presence of animals, but not their absolute number. This definition is inadequate, in that it makes no statement about the relationship between the count C and the unknown population size N. To be an adequate reflection of N, C must have some consistent relationship with N. This relationship is sometimes defined by noting that C must be positively correlated with N. For an index C to be useful, however, C must be a reasonable surrogate for N, both in hypothesis tests and in its association with covariates.

Consider the count of birds captured (or recaptured) at a mist-netting site as a possible index to the population size. The relationship of captured birds at a mist-net site to the actual population size can be expressed as

$$\mathbf{E}(C|p,N) = pN$$

where $\mathbf{E}(C|p,N)$ denotes the expected value of C conditional on the actual population size N, and p is the

proportion of animals encountered. In general, if p is not related to N, and is not 0, then C is a reasonable index of N. However, the correlation between C and N will depend on the variation of p, and any analysis of count data relies on some assumptions about either the magnitude of p or its consistency over any comparisons of populations that use counts. This has led to two major philosophical approaches to the analysis of index data.

Proponents of the first approach have said that "Using just the count of birds detected (per unit effort) as an index [of] abundance is neither scientifically sound or reliable" (Burnham 1981:324), and that "It is imperative in designing the preliminary survey to build in the capability...of testing homogeneity of the proportionality factor values..." (Skalski and Robson 1992:29). To apply this approach, an experimenter explicitly estimates p and tests for differences in p that can be confounded with the comparison of interest. For mist-nets, capture–recapture methods are used to estimate p (Kendall et al. *this volume*). If no differences in p are found, then the indices are used in analyses. However, without estimating p as a routine component of a study, these tests cannot be conducted, and the study will have little credibility (a point forcefully made by Anderson 2001).

In the second approach, indices are used in analyses without estimation of p. Instead, it is assumed that standardization and covariate analysis can be used to control variation in p that might invalidate hypothesis tests (e.g., differences in p might be confounded with treatments). Proponents of the second approach feel that it is impossible to design extensive studies to estimate p due to the practical constraints of low recapture rates and small sample sizes for most species in mist-net studies. In fact, many large-scale monitoring programs (such as the North American Breeding Bird Survey [BBS], Peterjohn and Sauer 1993) do not allow for estimation of p.

The first approach (in which p is estimated) should be considered in design of any field study, and the ornithological community increasingly attempts to estimate detectability in studies that count birds (e.g., Rosenstock et al. 2002). However, mist-netting samples are often too small to allow proper estimation, or the hypothesis tests based on the data have too low power to ever be able to test whether detection probabilities differ. In practice, many analyses are conducted on unadjusted counts of captured (or observed) birds.

ALTERNATIVE ESTIMATES OF POPULATION SIZE

Three distinct quantities are commonly referred to as the population size: first, N, the parameter (found only by censusing, which is almost never accomplished in bird monitoring); second, $\hat{N}$, the capture-recapture estimate, found by estimating p and defining

$$\hat{N}_i = \frac{C_i}{\hat{p}_i},$$

(Lancia et al. 1994); and third, C, the index. To investigators, it is not always clear how these quantities differ, and when it is appropriate to use $\hat{N}$ or C as a surrogate for N in hypothesis tests. To understand the consequences of this choice, we must consider two characteristics of the estimates, bias and precision.

Bias

The bias of an estimate is the difference between the expected value of the estimate and the parameter. For the capture–recapture estimate, the expected value of $\hat{N}$ is $\boldsymbol{E}(\hat{N}|N) \approx N$ (the estimator is slightly biased; Skalski and Robson 1992). In contrast, the bias of C is $\boldsymbol{E}(C|N) - N = pN - N = N(p - 1)$; hence C is always biased unless $p \equiv 1$.

Bias can be an extremely serious deficiency in an estimator, if it is not taken into account in hypothesis tests. The possibility that bias can differ among treatments should be considered in any hypothesis test that uses counts, and obviously invalidates use of the index as an estimate of population size. An additional consequence of the bias in C is that comparative tests of population size based on the counts may also be invalid. For example, suppose that we have replicate counts from sites 1 and 2. We are interested in testing a null hypothesis:

$$H_o: N_1 = N_2,$$

by comparing mean counts. Counts should only be used in this analysis if $p_1 = p_2$. Of course, this condition of equal p's is also necessary for any comparative test (e.g., a ratio analysis of productivity, where groups 1 and 2 would denote different age classes).

Bias is therefore a critical consideration for any analysis of count data. Unfortunately, after counts are collected, most statistical tests do not directly include an assessment of possible bias, so investigators do not become aware of these difficulties in the analysis.

Precision

At a single site, sampling error is the variance of the estimate conditional on the population parameter. Sampling error for a population estimate $\hat{N}$ is denoted by $V(\hat{N}|N)$. In a capture–recapture study, $V(\hat{N}|N)$ is estimated by assuming N and p are unknown but fixed, and estimating p from observed counts of marked and unmarked animals (Skalski and Robson 1992). If multiple sites are sampled, an additional factor, the among-site variance $V(N)$, is also a component of error, and the variance calculated among site estimates i, $V(\hat{N})$, is

$$\mathrm{V}(\hat{N}) = \mathrm{V}(N) + \mathbf{E}(\mathrm{V}(\hat{N}_i|N_i)),$$

where $\mathbf{E}(V(\hat{N}_i|N_i))$ is the expected value (average) of the within-site sampling errors. In most studies, $V(N)$ is the variance component of interest (Skalski and Robson 1992, Link and Nichols 1994).

If only counts are collected, this partitioning of sampling error and among-site variance cannot be conducted unless p is assumed fixed among sites, and known (Skalski and Robson 1992). Consequently, estimation of p is essential for studies in which estimation of variance components are of interest. Unfortunately, most studies of temporal variation in bird populations do not do this, and may provide incorrect results (Link and Nichols 1994).

Estimation of p still allows for use of C in hypothesis tests when p does not differ among populations to be compared. Skalski and Robson (1992) note that, unless $p = 1$, coefficients of variation of C will be smaller than coefficients of variation of $\hat{N}$ for a site. Hence, use of C in hypothesis tests will lead to higher power relative to tests based on $\hat{N}$, but only when p can be documented to be constant. Of course, if p is not constant the increased precision will only lead to an increased chance of a false rejection of the null hypothesis.

DEVELOPING A STRUCTURE FOR ANALYSIS OF COUNT DATA

The foregoing discussion provides a general view of the statistical properties of indices and capture–recapture-based estimates. However, investigators need specific methods for evaluation of the performance of indices and adjusted counts. Capture–recapture models provide a convenient framework for this evaluation. We can develop models for sampling the population, and see how counts and capture–recapture estimates differ in the context of the models. We provide an example of this based on the Lincoln index, as defined by Skalski and Robson (1992:63–64).

In the model, banding occurs in two periods (j = 1,2) at a single site. We use this notation:

N = number of animals
n_j = number of animals captured in period j
p_j = probability of capture in period j
q_j = 1 - probability of capture in period j
m = number of marked animals from period 1 recaptured in period 2.
$C = n_1 + n_2 - m$ = number of distinct captures

Under this model, the estimates of population size are

$$\hat{N} = \frac{n_1 n_2}{m} \approx \frac{(n_1+1)(n_2+1)}{(m+1)} - 1,$$

with sampling variance

$$\hat{\mathrm{V}}(\hat{N} \mid N) \approx \frac{N q_1 q_2}{p_1 p_2}.$$

The number of distinct animals counted is

$$C = n_1 + n_2 - m$$

with mean and variance

$$\mathrm{E}(C \mid N) = N(1 - q_1 q_2)$$
$$\mathrm{V}(C \mid N) = N q_1 q_2 (1 - q_1 q_2).$$

Under this model, we can directly estimate the bias and precision of counts and the capture–recapture population estimates.

Suppose that there are two sites, and a Lincoln experiment has been done on each. To test the null hypothesis that

$$\mathrm{H}_o: \mathrm{N}_1 = \mathrm{N}_2,$$

two alternative statistics can be used. The first is based on the capture–recapture-based estimate, using the statistic

$$z_{\hat{N}} = \frac{\hat{N}_1 \bullet \hat{N}_2}{\sqrt{\mathrm{V}(\hat{N}_1|N_1) + \mathrm{V}(\hat{N}_2|N_2)}}$$

The second is based on the counts of animals captured, using the statistic

$$z_C = \frac{C_1 \bullet C_2}{\sqrt{\mathrm{V}(C_1|N_1) + V(C_2|N_2)}}$$

Note that $z_{\hat{N}}$ and z_C do not test the same hypothesis. For $z_{\hat{N}}$, the null hypothesis is: H_o: $N_1 = N_2$, but for z_C it is: H_o: $\mu_{C1} = \mu_{C2}$ (where μ_{Ci} = mean count for i). These hypotheses are only the same when $p_1 \equiv p_2$.

To show the consequences of using z_C as a surrogate for $z_{\hat{N}}$, use the expected values given above in formulas for the z statistics, setting $N = N_1 = N_2$, $p_1 = p_{11} = p_{12}$, $p_2 = p_{21} = p_{22}$, and $p_1 \neq p_2$, to simplify the discussion. We can assess the differences in the tests for differing values of p_1 and p_2. For $z_{\hat{N}}$,

$$E(z_{\hat{N}}) = \frac{N - N(\pm \text{small bias})}{\sqrt{N\left(\frac{q_1^2}{p_1^2} + \frac{q_2^2}{p_2^2}\right)}} \approx 0;$$

and for z_C,

$$E(z_C) = \frac{N(q_2^2 - q_1^2)}{\sqrt{N(q_1^2 + q_2^2 - q_1^4 - q_2^4)}}.$$

In other words, $\mathbf{E}(z) \neq 0$ for a z statistic based on the C's, but $\mathbf{E}(z) = 0$ for the statistic based on the N's, thus tests based on z_C will have an inflated probability of a type I error rate (α) level. Using the expected values, we can quantify the inflation for a fixed N, p_1, and p_2 as

$$\alpha_{N,p1p2} = \overline{\Phi}[z_{\alpha/2} - E(z_C)] + \Phi[-z_{\alpha/2} - E(z_C)]$$

where Φ signifies the cumulative normal probability, and $\overline{\Phi} = 1 - \Phi$. Calculating these as a function of N with $\alpha = 0.05$, it is evident that the inflation of α increases both as a function of N, p_1, and p_2 (Table 1). When the total population size is moderately large (e.g., $N > 100$), the inflation in α is quite large for even small (5%) changes in p.

We conclude that minor changes in p between treatments can lead to large increases in type I error rates. When hypothesis tests are based on counts, differences in detection rates are confounded with differences in the actual population sizes; significant differences found in the test of equality of counts between populations may be entirely due to differences in p. Changes in p do not appear anywhere in the count-based analysis, and would be interpreted as rejections of null hypotheses by the investigator.

The changes in detection probabilities affect all aspects of hypothesis testing. For example, power (the probability of rejecting a "false" null hypothesis) is a function of the difference between the estimate and a hypothesized value of the parameter, and increases as the variance of the estimate decreases. Because variances decrease as sample sizes increase, test power increases with sample size. Consequently, increasing the observed power of a test when the estimate is biased leads to greater probability of error. Standard sample allocation procedures are therefore invalid, and lead to higher than nominal type I error rates.

A MORE GENERAL CASE

Suppose we have a study that only collects count data from $j = 2$ treatments, where C_{i1}, $i = 1,\ldots,I$, and C_{l2}, $l = 1,\ldots,L$ represent the counts for I replicate sites in treatment 1 and L replicates in treatment 2. Further, assume that for each treatment the counts are indices to population size, and that $p_1 \neq p_2$ (i.e., the detection probability is constant within treatments but differs between treatments).

To test whether H_o: $N_1 = N_2$, we use

$$z_C = \frac{\overline{C}_1 - \overline{C}_2}{\sqrt{\hat{V}(\overline{C}_1) + \hat{V}(\overline{C}_2)}}$$

which actually tests H_o: $\mu_{C1} = \mu_{C2}$.

The numerator of the test has expected value

$$p_1 N_1 - p_2 N_2$$

which, when the null hypothesis is true, equals

$$N(p_1 - p_2).$$

TABLE 1. THE ACTUAL ALPHA (α') ASSOCIATED WITH HYPOTHESIS TESTS ON COUNT DATA WHEN THE PROPORTION OF ANIMALS DETECTED CHANGES, FOR A FIXED TOTAL POPULATION SIZE

N	$\alpha'(\Delta p = 0.5–0.55)$	$\alpha'(\Delta p = 0.5–0.6)$	$\alpha'(\Delta p = 0.4–0.6)$
10	0.0574	0.0793	0.1820
50	0.0878	0.2020	0.6486
100	0.1267	0.3545	0.9117
150	0.1663	0.4932	0.9819
200	0.2063	0.6116	0.9968

We can use the argument given above to demonstrate the effect of differences in p between treatments on the hypothesis tests. Specifically, for any observed difference in counts (C_1 - C_2), the numerator of the test, we can ask whether, given that the mean population is of size N at both sites, what differences in proportion detected between treatments (denoted by Δp) would be expected to produce the observed z value.

For fixed N between treatments, Δp is

$$\Delta p = \left(\frac{E(C_1)}{N} - \frac{E(C_2)}{N} \right)$$

If these Δp values are small, the tests have little credibility. For example, Hanowski et al. (1993) presented data on mean counts of Downy Woodpeckers (*Picoides pubescens*) on two treatments, each based on 40 point-count sites. The estimates for the two treatments were 0.35 ± 0.09 (SE) and 0.17 ± 0.08. For fixed values of N, we calculate values of Δp that would produce the observed difference in means, given that both treatments have the same N ($N_1 = N_2 = N$). For example, if N equals 1.0 in both treatments, a Δp of 0.18 would be needed to produce the observed difference in counts, but if $N = 2.0$, a Δp of 0.09 will produce the observed difference in counts. If the counts are similar in magnitude to the actual population size (e.g., p is close to 1.0), then it is unlikely that changes in p are causing the observed differences in counts. However, if the p_m is much less than 1.0 (i.e., N is much greater than C), then relatively small differences in proportions detected between treatments will explain the differences between the observed counts. In this case, and in any analysis involving counts as surrogates for population size, it is informative to play "what if" games to evaluate whether the analysis is likely to be affected by differences in detection probabilities between treatments. To do this, postulate the detection probabilities and evaluate the consequences for the analysis. A similar procedure can be developed for any hypothesis test based on counts, such as testing for change over time or for ratios of counts.

CONCLUSIONS

In this paper, we have provided a framework for the analysis of count data, and identified some of the fundamental attributes of counts of birds captured as surrogates of population parameters.

- Counts are always biased unless $p = 1$. This means that counts do not estimate population size, but estimate population size times p.
- Counts are always more precise than adjusted population estimates. This is due to the bias in the estimate ($p < 1$), and the additional error associated with estimating p that occurs in the adjusted estimates. Counts are most precise when $p = 0$, which demonstrates that the increased precision of counts is not useful for hypothesis testing unless differences in p are accommodated in the analysis.
- Sample allocations based on C are not appropriate, because increased samples lead to more precise estimates of $\mathbf{E}(C)$ rather than of N. This amplifies the bias in statistical tests.
- Simple analyses of C omit discussion of bias. Hypothesis tests do not accommodate the possibility of differences in p, and will produce inflated α levels with even moderate differences in p.
- We can use mark–recapture structure to incorporate bias into the analysis, and simulate the effects of changes in p between treatments. If no estimate of p is available, we can model possible effects of variation in p on analysis.
- It is wrong to eliminate p from analyses of count data. The best way of incorporating p in the analysis is to estimate p for each treatment, test for differences in p between treatments, and if necessary incorporate the ps in the hypothesis tests (e.g., Skalski and Robson 1992). If p cannot be estimated, then it must be demonstrated that the hypothesis test is likely to be valid for moderate differences in p between treatments. However, ignoring the possibility of differences in p will lead to analyses with low credibility.