ESTIMATING CLOSED POPULATION SIZE AND NUMBER OF MARKED ANIMALS FROM SIGHTING DATA

A. NEIL ARNASON, Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

CARL J. SCHWARZ, Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

JON M. GERRARD, Manitoba Institute of Cell Biology, 100 Olivia Street, Winnipeg, MB R3E 0V9, Canada

Abstract: We describe a new estimator of population size that can be formed when independent sightings are made of marked and unmarked animals in a closed population where a subset of the population is individually marked. Each marked animal must bear a unique mark but the number of marked animals alive in the population is unknown. The estimate can be used when no recaptures or removals of animals are possible during the experiment. An example is estimating the number of immature bald eagles (*Haliaeetus leucocephalus*) on a lake some years after banding of nestlings. We derive the maximum likelihood estimates for population size and number of marks, and we show how to develop confidence intervals and perform goodness-of-fit tests. Criteria are developed for determining the number of sightings required to yield satisfactory estimates.

We derived a new estimator of population size that makes use of data from animal marking and sighting experiments. The estimator applies to experiments in which an unknown number of uniquely marked animals exists in the population at the time when the population is surveyed. The population is surveyed multiple times and is assumed to be closed throughout the time of the surveys; that is, it is not subject to any significant additions or losses during this period. During the surveys, sightings are made by spotting or locating animals and recording whether the animal is marked or unmarked and, if marked, its identity. Animals are neither captured nor removed during the surveys, so no marking of unmarked animals takes place at this time. Sightings can be made with no physical handling of the animal. The essential feature of a sighting is that the marked status and identity of the animal can be determined but, unlike when sampling or capturing animals, the marked status cannot change, and the animal cannot be removed. At the end of the survey period, the experimenter knows how many different marked animals were sighted and how often each was sighted. The experimenter also knows the total number of unmarked sightings but not how many different animals this represents nor how many marked or unmarked animals were never seen. There are 2 unknown quantities to be estimated: N, the total population size at the time of the surveys and M, the number of marked animals in the population. That M is unknown is what makes this estimator distinct from any existing estimator of closed population size.

The experiment permits considerable flexibility in how the M marked animals were established in the population prior to the survey. Animals may have been captured, marked, and released at a single point in time prior to the survey or over an extended period of time; the marked subset can be formed by the addition of new, marked animals, prior to the survey. Marked animals could then be lost through emigration, death, or tag loss prior to the onset of the survey period. Any type of marking method that permits unique identification of the animal at sighting, including radio telemetry or self-markings where animals are identifiable from individual physical characteristics, may be used . An important assumption of our method is that

marks are read accurately and that marked and unmarked animals are equally likely to be

sighted, so the choice of mark and marking method must be made carefully.

A boat survey of bald eagles on Besnard Lake, Saskatchewan, carried out over several years

to monitor both adult and immature numbers of eagles (Gerrard et al. 1990) motivated the

development of our estimator. In 1973-75, 56 nestling eagles were color-marked with patagial

tags. During 16-30 July 1977, sightings of immature eagles were made daily on an opportunistic

basis while the experimenters were conducting a nest survey. On most days, only a few (1-3)

immature eagles were sighted per day. On 2 occasions however, 24-27 July and 25-29 July, a

more systematic and intense census of the entire lake was made. A census involved an

examination of every second 8-km section (out of 50 sections on the lake in total) of shoreline by

observers traveling in a boat at 8-16 km/hr, approximately 100 m from the shoreline. The 2

censuses produced 36 and 23 sightings, respectively. Surveys were done when weather was calm

and visibility was good. Each day, the observers recorded the number of unmarked birds and the

identities of all marked birds seen.   Throughout this period, immatures frequently occurred in

groups, and the numbers per sector were clumped. However, immature birds did not tend to be

seen repeatedly in the same sections, but were highly mobile and moved around the lake rather

than occupying particular areas or territories. Although there are neighboring lakes also

populated with eagles, the birds on Besnard Lake probably spent little, if any, time off the lake

during the survey period. The census data provides a direct estimate of population size (Gerrard

et al. 1990) which is a helpful check on the estimate we developed.

Related Methods:—Experiments that involve marking and sightings have been called tag-

resighting (Brownie and Robson 1983), mark-resight (Brownie 1987), or capture-resight

(Minta and Mangel 1989) experiments. These terms most appropriately apply to the

generalization of the Jolly-Seber (Seber 1982:196-217) open-population mark-recapture

experiment in which some samples or intersample periods involve sightings only of marked

animals; unmarked sightings are either unrecorded or ignored. All other sample times involve

capture of animals, recording of marked and unmarked, and the application of marks to the

unmarked animals before all, with the exception of losses-on-capture, are returned to the population. This process, whereby all unmarked animals in a sample are removed by loss-on-capture or are marked before being released, is called continuous marking by Minta and Mangel (1989). If continuous marking takes place, all sightings are re-sightings. The experiment described by Minta and Mangel (1989) does not involve continuous marking; it is identical to the one that we propose with the crucial distinction that they assume that M is known. In both cases, the unmarked sightings may involve animals sighted for the first time so we suggest the term marking and sighting experiments.

The mark-recapture alternatives to marking and sighting experiments for closed populations are the Petersen method and its extension to multiple samples with continuous marking, i.e., the Schnabel method (Otis et al. 1978). The Petersen estimate, in its simplest form, applies to a 2-sample experiment where M animals out of N are marked in a first sample and, in a second sample of n distinct animals, m are found to be marked. The single unknown, N, can then be estimated from the Petersen ratio equation:

$$\frac{M}{N} = \frac{m}{n} \tag{1}$$

The Petersen equation also applies to the marking and sighting experiments of Minta and Mangel (1989) where M is known. In this case, the second "sample" is the aggregate of all surveys, and m is the total number of sightings of marked animals out of n sightings. We show (Appendix) that, provided the assumptions of the Petersen method are met, the information on the number of distinctly marked animals seen and on the frequency of sightings of the marked animals, can be ignored for purposes of estimating N, but is still useful for testing assumptions. The crucial assumption for an unbiased simple Petersen estimate is that inclusion of an animal in the second sample is independent of its inclusion in the first (Seber 1970). Minta and Mangel (1989) give a good discussion of what this assumption implies for conducting sighting experiments. To form a standard error or confidence interval for N, all animals in the second sample are assumed to be independently and equally catchable. In extending the Petersen estimate to marking and sighting experiments, these assumptions must apply at every sighting survey.

Minta and Mangel (1989) make a detailed argument as to when and why marking and sighting experiments are to be preferred over 2-sample Petersen and multi-sample, continuous-marking experiments. They argue that sightings are cheaper and easier than recaptures; that the assumption of independence between capture probability of the initial marking and subsequent resighting, although still critical, is more likely to be satisfied; that sightings permit estimation from continuous surveys involving different degrees of effort without artificially grouping them into sample times. Minta and Mangel (1989) then develop a Monte Carlo method that permits them to relax the assumption of equal catchability (but not the independence assumptions) for purposes of estimating the confidence interval for N. When surveys are regular, involve equal sighting effort, and yield reasonable numbers of sightings per survey, the Petersen ratio can be applied to each survey and the estimates combined to form a single estimate of N (Bartmann et al. 1987). These appear to be the only methods for use with sighting data from closed populations, and both were designed for use with radio telemetry where it is reasonable to assume and possible to confirm that M, the number of transmitter-equipped animals, is known.

Several studies from the literature on mark-recapture are relevant to our study. Otis et al. (1978) gave a thorough treatment of continuous-marking methods for closed populations, including the Schnabel model and extensions that relax its assumptions of equal and independent catchability. Darroch and Ratcliff (1980) and Sprott (1981) have studied the Schnabel estimate with continuous sampling; that is, where every capture is considered a sample of size 1. This is the continuous-marking analog of our sighting experiment. Chao (1988, 1989) has extended the work of Otis et al (1978) on methods that are robust to unequal catchability.

The problem of estimating population size from sighting data is statistically similar to problems in estimating the number of different classes in a population. These problems arise when trying to estimate the number of species in a community, or the number of genes on a chromosome. In our case, the number of classes is the number of uniquely marked animals. Lewontin and Prout (1956) discussed the case where classes are equi-probable. The work of Lewontin and Prout (1956) applies to situations in which $N = M$ because all animals are uniquely

identifiable, so there are no unmarked sightings. Our problem is to extend the result of Lewontin and Prout (1956) to account for the *n - m* unmarked sightings which then permits estimation of *N* and *M*.

THE ESTIMATOR AND ITS PROPERTIES

Definitions.—The following summarizes the notation and assumptions for the method. Later, we will discuss the extent to which the assumptions can be relaxed.

1. The population is made up of a fixed but unknown number (*N*) of animals, and the population is closed to changes in size (through immigration, emigration, recruitment, and mortality). The population contains a fixed but unknown number (*M*) of marked animals, each of which is uniquely identifiable, without error, if it is sighted.

2. A survey is carried out in which a total of n independent sightings of animals occur. We assume that, at each sighting, every animal has an equal probability of being sighted and that each sighting is independent.

3. The data obtained from the surveys is the total number of sightings of unmarked animals, and the frequency of sightings of the marked animals.

   We will use the following notation:

f_i = number of marked animals seen i times, where $i = 1, 2, ...$; note that no f_i nor i itself can exceed *n* and that f_0, the number of marked animals that are never sighted, is unknown.

m' = total number of different marked animals seen, so $m' = \sum_{i>0} f_i = M - f_0$.

m = total number of sightings of marked animals, so m = $\sum_{i>0} if_i$.

There will also be multiple sightings of animals among the *n - m* unmarked sightings, and presumably these will occur with similar frequencies as among the marked animals, but these

frequencies are unknown.　Although estimates of N are of primary interest in the case of our eagle data, there may be situations when estimates of M are of equal interest because they permit estimation of a survival rate. If all the marks were released t time units prior to the onset of the surveys and the durations of the marking period and of the survey period are short relative to the elapsed time t, then a survival rate could be estimated. If the number of marked releases M_0 is known, then the survival rate per t time units is $\phi = M / M_0$. We will show how to estimate both M and N and their confidence intervals.

Statistical Methods. –– Estimates are obtained by the method of maximum likelihood: that is we write down the probability of what was seen ($n - m$ and the f_i) as a function of the sample size (n) and the unknown parameters (N, M). The maximum likelihood estimates, $\hat{N}$ and $\hat{M}$, are the values that make this likelihood function, denoted as L(N, M), as large as possible. Further discussion of maximum likelihood, as applied to mark-recapture models, can be found in Appendix A of Otis et al. (1978).

There are several practical problems that arise in the use and interpretation of the estimates: when are the data useable to form estimates, how to find the maximum likelihood estimates, how to deal with the constraint that arises from knowing that M must be at least as large as the number of different marked animals sighted (m'), how to form confidence intervals so that they give proper coverage of the true values, when do the estimates give adequate precision, and how to form tests of the assumptions of the model. Many of these problems are most easily investigated by using a computer to simulate the sighting experiment in a population of N animals with M of them marked. We randomly selected n animals out of N with replacement in such a way that the assumptions of independence and equal sightability were satisfied. The computer program kept track of which animals were marked, their sighting frequencies (f_i), and the number of unmarked sightings, and used them to form the estimates, standard errors, and confidence intervals. The properties of these quantities, for given $n, N,$ and M can then be studied by replicating this sampling experiment hundreds of times.

Computing Estimates. –– There is no explicit expression for the maximum likelihood estimates $\hat{N}$ and $\hat{M}$ but we show, in the Appendix, that the unconstrained estimates satisfy

$$\frac{\hat{M}}{\hat{N}} = \frac{m}{n},$$ (2)

(the Petersen ratio, except that m and n can both include multiply sighted animals), and

$$\frac{m}{\hat{M}} = \left[\frac{1}{\hat{M}} + \frac{1}{\hat{M} - 1} + \dots \frac{1}{\hat{M} - (m' - 1)} \right].$$ (3)

The individual f_i do not appear in these 2 equations and are not needed to form the maximum likelihood estimates. We need only the 2 statistics computed from the f_i: m and m'. In statistical estimation theory, m and m' are said to be sufficient for inference about N and M, given n. Equation 3 defines the maximum likelihood estimate for M without reference to N and may be solved first by trial and error or computer search. The solution, however, may violate the constraint $\hat{M} \geq m'$. There may also be a constraint on the highest possible value for $\hat{M}$ if we know the total number of marks that were ever released into the population (M_0). When the maximum likelihood estimate violates either constraint, it is said to be inadmissible. The usual practice in mark-recapture studies is simply to adjust the estimate to the nearest integer within the constraint boundaries (Otis et al. 1978, Appendix A). However, we used the unadjusted $\hat{M}$ to compute $\hat{N}$ (eq. 2) and the standard errors and confidence intervals because our simulation study showed that use of the adjusted $\hat{M}$ introduces substantial bias in $\hat{N}$, its standard error, and the confidence interval coverage for N and can cause computation problems such as zero divides and negative variances. After evaluating the estimates associated with N, $\hat{M}$ and its confidence interval can be adjusted. Thus, even though M is an integer and must be greater than m', $\hat{M}$ should be estimated to ≥ 2 decimal places and without constraint: there are 4 possible cases:

(1) $m' = 0$, $m = 0$. No marked animals are sighted. The experiment is uninformative and no estimates can be formed.

(2) $m' = 1$, $m > 1$. A single marked animal is sighted several times. The likelihood L(N,M) is unbounded at $N = 0$. This means that the unconstrained maximum likelihood estimates and their standard errors cannot be formed. The adjusted estimate is $\hat{M} = m' = 1$

which, in equation 2, gives $\hat{N} = n/m$, but no standard error or confidence interval can be formed for this adjusted $\hat{N}$.

(3)   m′ = m, $m \geq 1$. No marked animal is sighted more than a single time. This can happen only if $m \leq M$. $\hat{N}$ is infinite and, as in case (2), no standard error or confidence interval can be formed.

(4)   In all cases other than (1)-(3), equation 3 has a unique solution with $\hat{M} > m′- 1$.   When $m′- 1 < \hat{M} < m′,$ $\hat{M}$ is inadmissible, but the inadmissible value is used to find $\hat{N}$ using equation 2 and then the standard errors and confidence intervals before adjusting $\hat{N}$ and $\hat{M}$ to $m′$ and $n\,m′/m$, respectively.

The variance-covariance matrix for $\hat{N}$ and $\hat{M}$ is formed from the inverse of the actual information matrix (Appendix) and gives the following standard error formulae:

$$\hat{S}(\hat{N}) = [(A + B) / D]^{1/2},$$

and $$\hat{S}(\hat{M}) = [(A - C) / D]^{1/2},$$

where $$D = A \cdot B - C(A+B)$$

and $$A = (n-m) / (\hat{N} - \hat{M})^2,$$

$$B = \left[\frac{1}{(\hat{M})^2} + \frac{1}{(\hat{M}-1)^2} + \ldots \frac{1}{(\hat{M}-(m′-1))^2}\right],$$

and $$C = n / \hat{N}^2.$$

The covariance between $\hat{N}$ and $\hat{M}$ is A / D. Our simulations showed that the usual method of constructing a 95% confidence interval for N as $\hat{N} \pm 1.96\,\hat{S}(\hat{N})$ gives poor coverage; that is, intervals formed in this way actually cover the true N in far fewer than 95% of the replications. This results from small sample bias, correlation in $\hat{N}$ and $\hat{S}(\hat{N})$, and non-normality in the distribution of $\hat{N}$. All are problems that frequently affect mark-recapture estimates (Manly 1984) and which can be dealt with by transformation methods (Otis et al. 1978, Sprott 1981), by profile relative likelihoods (Morgan and Freeman 1989), and by data intensive techniques such as Monte Carlo (Minta and Mangel 1989), jackknife, and bootstrapping methods. We did not investigate the latter methods because we found the simpler methods

worked adequately. Otis et al. (1978) investigated the effect of a number of transformations on confidence interval coverage of *N* and found none of them worked very well. However, Sprott (1981) showed analytically that the most appropriate transformation for Schnabel-type estimates was one which was not investigated by Otis et al. (1978). Because of the close resemblance of our estimator to the one investigated by Sprott, we were not surprised to find that his inverse cube root transformation greatly improved confidence interval coverage for N. The 95% confidence interval, (N_L, N_U), will not be symmetric about $\hat{N}$. It is formed as follows:

$$\hat{T} = \hat{N}^{-1/3}$$

$$\hat{S}(\hat{T}) = \hat{T} \cdot \hat{S}(\hat{N}) / (3 \hat{N})$$

$$(T_L, T_U) = \hat{T} \pm 1.96 \, \hat{S}(\hat{T})$$

$$(N_L, N_U) = (1/T_U^3, \ 1/T_L^3)$$

The transformation is effective at improving the approximation to normality even for very small samples and numbers of marks, and its effectiveness improves rapidly with larger values of *M* (Fig. 1).

  Interval estimates for *M* can be formed by the method of profile relative likelihoods. The method is outlined in the Appendix and is implemented in our computer program. The method produces a 95% profile confidence interval (M_L, M_U) by finding the extreme extents in the *M* direction of the 15% contour of the relative likelihood function $L(N,M)/L(\hat{N}, \hat{M})$ (Fig. 2). If necessary, $\hat{M}$, M_L, and M_U should all be adjusted to admissible values and can then be converted to the corresponding values for survival, ϕ, by dividing through by M_0. We did not examine the coverage of the interval because we were not interested in estimating *M* or the survival rate of marked eagles. However, the method has been found effective in obtaining interval estimates for survival parameters in band-recovery models (Morgan and Freeman 1989) and generally produces asymmetric intervals that allow for non-normality in the distribution of such estimates. Profile intervals are difficult to compute for *N* in this model although they can easily be determined approximately by visually assessing the extents of the 15% contour in the *N*

direction. In the few situations where these contours were plotted (Fig. 2), the profile interval corresponds almost exactly to the confidence interval formed by the transformation method.

Survey Design Criteria.—We used the simulation program to examine the effects of a range of small sample situations, i.e., <200 sightings in populations of up to 100 animals and <12 marked animals. These ranges centered approximately on those for our eagle population. The summary of the simulation in Table 1 was designed to quantify the likelihood of obtaining the uninformative estimates described above in cases (1)-(3), or inadmissibly low estimates of M as in case (4). It was also designed to assess the contribution made by bias in $\hat{N}$ to poor coverage of the confidence interval. This can be assessed by examining the effective bias (Arnason and Mills 1981) measured by $100 \cdot (AV[\hat{N}] - N) / SD[\hat{N}]$ where AV and SD denote the mean and standard deviation of $\hat{N}$ over the replications of the experiment . If the effective bias is less than 50%, then the contribution of small sample bias to lack of coverage is negligible. The coverage can be measured directly. It is also affected by non-normality in the (transformed) estimate. Reasonable criteria that are likely to yield acceptable estimates are: (1) the "bad" m, m' values described in cases (1)-(3) above occur less than 10% of the time; (2) the effective bias in $\hat{N}$ is less than 50%, and (3) the estimated coverage does not differ significantly from 90% or greater. Acceptable estimates result whenever the expected value of $m', E(m')$, exceeds 5, and the expected value of m, $E(m)$, exceeds $2E(m')$ (Table 1). For planning purposes, one can determine what number of sightings (n) will be needed to produce good estimates given (guessed values of) N, M because

$$E(m) = nM/N ,$$

and
$$E(m') \approx M(1 - e^{-n/N}) . \tag{4}$$

Equation 4 is derived from equation (16) of Lewontin and Prout (1956) and the Petersen ratio given by our equation 1.

The simulations can also tell what precision to expect for different expected values of the sufficient statistics. Precision is commonly judged by the coefficient of variation of the estimate (Table 1), but this can be misleading if the estimate is non-normally distributed or if the

estimated standard error of the estimate is biased. This latter bias can be judged by comparing the standard deviation of the estimate, SD[$\hat{N}$], with the average of the standard error, AV[$\hat{S}(\hat{N})$]. For the acceptable estimates in Table 1, the standard error underestimates the true value by 5-20%.   Where bias is this large and non-normality is extreme, a better measure of precision is the average quarter width of the confidence interval. Expressed as a percentage of the estimate (Table 1), it can be compared to the percent coefficient of variation. If the estimate is normally distributed, the 2 statistics would be much the same. The increase in percent precision relative to the coefficient of variation reflects the degree of non-normality and its cost in precision.

   Testing Assumptions.—The entire discussion of the estimates to this point assumes that the assumptions of the model hold. It is important to use whatever data or biological knowledge is available to test the assumptions. If animals are sighted independently and with equal probability, then the sighting frequencies f_0 , f_1 , f_2 , ... should follow a binomial distribution on $\{0, 1,...n\}$ (or, to a very good approximation, a Poisson distribution) with mean equal to the mean sighting frequency. This is the basis for the test used by Minta and Mangel (1989) and is easily applied by them because they know f_0 , the number of marked animals that are never sighted, and they can estimate the mean of the distribution by the total number of marked sightings divided by M, the (known) total number of marked animals. In our situation, we do not know f_0 or the mean number of sightings per marked animal, but we do know the conditional mean number of sightings given that an animal is seen at least once (m/m'). We, therefore, carry out the test by comparing the zero-truncated Poisson distribution with this mean to the f_1 , f_2 , ... We show in the Appendix how to compute the estimated frequencies, $\hat{f}_i = m' P'_i$, and carry out the Chi-square test of goodness-of-fit. Minta and Mangel (1989) present several examples of the application of their test which show that it is good at detecting differences in sightability among the marked animals. When the test fails, it usually does so because some animals are sighted far more frequently than would be expected under the assumptions of independence and equal sightability. Their test can also detect situations where more marked animals go unsighted than

would be expected under the assumptions whereas ours cannot, but neither test can detect the most dangerous assumption failure: situations where the average sightability of marked and unmarked animals is different. No statistical test will ever detect this situation for we can never make comparisons with what is not observable − in this case, the sighting frequencies of the unmarked animals. The experimenter's biological judgement of the animal's behavior in relation to the method of marking and sighting must be used to assess if it is likely that this can occur or has occurred.

APPLICATION TO DATA

Here we apply the methods of the previous section to 3 data sets (Table 2). The first is the subadult eagle data of 16-31July 1977 reported by Gerrard et al. (1990). A survey of $n = 107$ sightings gave sufficient statistics $m = 7$ and $m' = 3$. These statistics were also the most frequent outcome for m and m' in the simulation experiment (Fig. 1a) where $n = 100$, $N = 50$, and $M = 3$. The second data set is hypothetical with $n = 100$, $m = 21$, and $m' = 7$. This was one of the more common outcomes of the simulation experiment with $n = 100$, $N = 50$, and $M = 12$ (Fig. 1c). Thus, both the eagle data and the hypothetical data might reasonably have arisen in a population of 50 birds, but the hypothetical data shows the effect of having 4 times as many marked birds in the population. The third data set is the badgers (*Taxidea taxus*) in Wyoming data from Minta and Mangel (1989) which we will use to compare with their analysis as an example of the effect of assuming M known or unknown. The data, estimates, and goodness-of-fit test for these data sets are shown in Table 2.

We plotted contour plots of the joint relative likelihood, $R(N, M) = L(N, M) / L(\hat{N}, \hat{M})$, and the profile relative likelihood, $R_P(M)$ for the first 2 data sets (Fig. 2). These plots permit us to see just what values of $\hat{N}$ and $\hat{M}$ are supported by the data and, by imposing the constraint regions on these plots, how to adjust inadmissible estimates.

Eagle Data.——The eagle data gives maximum likelihood estimates for (N, M) of (39.64, 2.59) but this lies in an inadmissible parameter region (Fig. 2a). The best point estimate is then $\hat{N} = 45.86$ which maximizes $R(N, M)$ along the constraint profile $M = m'$. This agrees well with

the direct sample estimates of 48 and 54 subadults obtained from the 2 intensive lake censuses (J. Gerrard, pers. commun.) and with the average July estimates (Gerrard et al. 1990) of 45.0 (standard error = 5.2). The 95% confidence interval for N formed from the inverse cube root transformation (18.5, 110.7) corresponds almost exactly to the 95% profile confidence interval: the bounds of the 0.15 contour in the *N*-axis direction (Fig. 2a). The coefficient of variation of the estimate is a little under 40%, which is optimistic compared to the 50% precision obtained as the quarter width of the confidence interval relative to the estimate. The profile relative likelihood for *M*, (Fig. 2b), also has its maximum at an inadmissible *M*. The most likely admissible *M* is 3 *(= m′)* and support for other *M* values drops off rapidly; *M* = 4 is about half as likely as *M* = 3, and the 95% profile confidence interval includes only values from 3 to 6. If all the eagles had been tagged at a single point in time so that a meaningful survival rate could be computed from $\hat{M}$, the estimate, $\hat{\phi} = \hat{M} / M_0$ would be remarkably precise. For example if M_0 = 50 birds had been released a year prior to the surveys, the annual survival rate would be estimated at 6% with a 95% confidence interval of 6.0 - 12.1. The goodness-of-fit test is not informative when there are only 2 observed sighting frequency classes. Overall, our estimates give barely acceptable results. They help to confirm estimates obtained from direct survey results (Gerrard et al. 1990) but are not precise enough to be very useful on their own. The lack of a goodness-of-fit test means that we must rely entirely on biological experience to be assured that the assumptions hold. As the simulations indicate (Fig.1a and the lines for n = 100, *N* = 50, and *M* = 3 in Table 1), we could not have expected much better with such a low number of marks in the population.

    Hypothetical Data.––With the higher number of marked animals in the hypothetical data set, however, the estimate (Table 2) improves substantially, to the point of being useful on its own. The point estimates, $(\hat{N}, \hat{M}) = (45.71, 9.60)$, are admissible, and the precision for $\hat{N}$ is under 25%. A much wider range of *M* values is supported by the profile relative likelihood (Fig. 2d) but the coefficient of variation of $\hat{M}$ (and hence of any ϕ formed from it) is nevertheless smaller (15%) than for the eagle data (25%). The goodness-of-fit test indicates no detectable problems

with the independence and equal-catchability assumptions. This is to be expected because we used these mechanisms to generate these simulated data.

  Badger Data.––The third data set arises from snowtracking of badgers in Wyoming (Minta and Mangel 1989). There were 15 radio-tagged badgers and, over a 2-month period, 107 sightings were made where a badger snow track could be followed to a terminal hole and the marked status and identity of the animal established. The frequencies with which animals were sighted (Table 2) add up to 13 ($= m'$), the number of marked animals sighted at least once. There were 2 badgers never encountered in the survey (f_0), and a total of 68 ($= m$) marked sightings. Minta and Mangel (1989) give a thorough discussion of the survey method and the possible sources of heterogeneity and unequal sightability. They conclude from this, and from the failure of their goodness-of-fit test, that there is some heterogeneity in sightability. The variance in the number of sightings per animal (10.25) is more than double the mean (4.53), indicating a clumped distribution. The Petersen estimate (adjusted for small-sample bias, as described by Minta and Mangel, 1989) for N is 23.5 with a 95% confidence interval of 20.2 - 26.8. Their Monte Carlo estimate produces the same point estimate because it fixes the mean sighting frequency at the observed value, but the clumped distribution produces a different confidence interval of 21- 28.

  Minta and Mangel (1989) do not explicitly state that they checked to see if the 2 unsighted badgers were present and their radio transmitters were functioning throughout the survey, although they do stress that such checks are important in general. By applying our method to these data, we can determine whether the data alone support the presence of all 15 animals. The first thing to note (Table 2) is that, if the 2 unsighted animals are ignored, the goodness-of-fit test under the hypothesis of equal, independent sightability is met. That is, the failure of the goodness-of-fit test in Minta and Mangel (1989) is almost entirely due to an unlikely value for f_0 . The range of possible values for M is doubly and narrowly constrained; it cannot be <13 or >15, but our model supports only 13 and 14 (after adjusting the confidence interval for M given in Table 2) with 13 being by far the more likely value. Thus, the data are perfectly consistent

with the hypothesis that there are only 13 (or 14) marked badgers available to be sighted, and they behave as if they were independently and equally sightable. Under these conditions, the estimate for N and its confidence interval bounds (Table 2) are all roughly 4 less than those for the Petersen estimate. We stress that these conclusions are derived entirely from the data, without taking account of information ancillary to the surveys that might shed light on the true value of M.

It is interesting to compare the results of the badger and eagle analyses. By coincidence, both experiments produced the same number of sightings overall (107). Our best estimate is that the badger population was around half that of the eagles (20 as opposed to 40) but had a much larger number of marked animals (13 as opposed to 3) and mean sighting frequency per animal (5.2 as opposed to 2.0). The precision is better for the badgers (7.5 as opposed to 50 percent), showing that even in very small populations it is possible to get good estimates.

DISCUSSION

The estimator of population size we propose has certain advantages over other methods of estimation based on marked recoveries or resightings. The main advantage is that it does not require the continuous capture or removal of animals, and it exploits the individual identifiability of animals as the result of previous marking or banding efforts. Indeed, the method might apply to some species where some of the animals are individually identifiable from natural markings, although if all animals are individually identifiable the methods of Darroch and Ratcliff (1980) or of Chao (1988) should be applied. Because there are no unmarked sightings in this case, it is equivalent to situations where every unmarked animal is marked or removed on capture; in both cases, the frequency of sighting of every different animal encountered is known. Another advantage of our method is that, unlike Schnabel-type methods (Otis et al. 1978), there is no need for well-defined sample periods which introduce additional nuisance parameters into the models (such as p_i, the probability an animal is captured in sample i) and additional operational problems (number, spacing, duration, and intensity of each sample). Further advantages of sighting experiments are discussed by Minta and Mangel (1989).

The disadvantages of this estimator are that the estimation method is operationally limited, and the estimate is sensitive to the assumptions of independent, equiprobable sightings in a closed population. By operationally limited, we mean that the experimenter has little control over *M* and so can increase precision only by increasing the number of sightings, n; however, because sightings are generally sequential, a larger n takes more time and at some point the validity of the closure assumptions would begin to be undermined. Moreover, because *M* does not increase as *n* increases, the gain in precision for a given increase in n is probably much less than for the Darroch and Ratcliff estimate. The Monte Carlo technique could take account of some degree of unequal catchability, but it would only account for unequal catchability among animals sighted at least once and, as with the method of Minta and Mangel (1989), would in no way account for bias arising from non-independence. Robust techniques like those of Chao (1988) might be developed for sighting experiments both with and without knowledge of *M*, but deriving an estimate and exploring its properties against a variety of models of assumption failure is a major task. In our opinion, these problems are best dealt with by careful survey design and execution.

The assumptions of equal-probability, independent sightings, and population closure are never likely to be met in a real population, but these assumptions do not have to hold absolutely. What is important is that, over the entire survey experiment, each animal has the same probability of being never sighted, sighted once, sighted twice, etc. (that is, all animals' sighting frequencies are identically distributed). This could happen even if animals were temporarily immigrating and emigrating or if, on 1 day, the survey was carried out in a restricted area so that some animals had no chance of being sighted on that day. Provided over *all* days each animal had an equal chance of being sighted, the estimate would not be seriously biased. Despite this, and even though the data over surveys is pooled for purposes of analysis, the assumption of independence of sightings over surveys is an important one. Surmounting this problem is the most difficult and important practical task facing biologists who wish to use sighting methods. Sightings must be randomized in space and time to minimize the chance of seeing some animals

preferentially to others. The movement rates and behavior of the animals also contribute to this randomization. One might be tempted to build up n in a cost efficient manner by conducting multiple surveys covering the same ground (e.g., a first survey going from A to B and a second survey on the return from B to A) on the same day. If sightability is largely determined by the location of an animal (e.g., for deer, animals grazing in the open have high sightability; animals under cover have low sightability) and there is little mixing of animals among locations between surveys, then bias is likely to result. For some highly mobile animals, this might not be a problem. Minta and Mangel (1989) give a great deal of helpful advice on this aspect of planning surveys along with several further practical examples.

Variability across animals in catchability or sightability can cause bias in $\hat{N}$. Otis et al. (1978) distinguish two types. Behavioral (b-type) heterogeneity, caused by a response to capture, can produce underestimates (trap-happiness) or overestimates (trap-shyness). Intrinsic (h-type) heterogeneity, that occurs when each animal's time-constant capture probability varies from animal to animal, will produce underestimates of N. These types also apply to the sighting experiment, although marked animals are typically captured only once, when the mark is first applied. Bias in $\hat{N}$ occurs if marked animals always, or even on average, are more sightable than unmarked. This is analogous to trap-happiness and leads to an $\hat{N}$ that seriously underestimates N.   This could happen for direct reasons (the mark itself makes the animal more visible) or indirect reasons (some behavioral or physical property of the animal made it easier to catch for marking purposes than other animals and also made it more likely to be sighted than other animals). Overestimates result if marked animals are less likely to be sighted than unmarked animals. Reduced sightability could result from "trap-shyness" if the survey period closely followed the marking period, or from visibility problems with the marks that cause some marked animals to be mistaken for unmarked. Therefore animals must be seen well enough that the investigator can reliably determine whether or not a marker was present. For example, we only included an eagle sighting in the total count (n) if we saw the bird well enough that we

would have seen a marker if it had been present. One advantage of using radio telemetry as a marking technique is that it reduces the subjectivity of such judgements.

Intrinsic (h-type) heterogeneity does not always produce underestimation as it does in the continuous-marking experiments of Otis et al. (1978). Unequal sightability that affects marked and unmarked animals equally will not bias $\hat{N}$. Otherwise, the bias will be equivalent to that caused by trap-happiness or trap-shyness, according as the average sightability of the marked animals is greater than, or less than, that of the unmarked animals.

Another source of bias is non-independence of sightings. Non-independence, in the form of sightings of flocks or other groupings does not introduce bias unless group size or some other group attribute affects sightability.   Non-independence that results from inadequate randomization between surveys, even if it causes no bias in $\hat{N}$ because it does not affect the differential sightability of marked and unmarked animals, will bias the standard error and confidence interval, generally making the estimator appear more precise than it really is. Consider the extreme example where every sighting is immediately repeated giving identical results. Here all the statistics are double what they should be, leading to higher precision than is warranted.

The estimator we developed works reasonably well in small populations with small marked fractions provided the number or marks (*M*) is >5 and the number of sightings, *n*, is >*N* (Table 1). For acceptable precision, n should be at least double *N* (Table 1). We did not investigate the properties of the estimate for larger values of *n, N*, and *M* but our computer program (written in FORTRAN with IMSL calls) is available from the senior author and can be used to investigate additional properties of the estimator.

LITERATURE CITED

ARNASON, A. N., AND K. H. MILLS. 1981. The bias and loss of precision due to tag loss in Jolly-Seber estimates for mark-recapture experiments. Can. J. Fish. Aquatic Sci. 38:1077-1095.

BARTMANN, R. M., G. C. WHITE, L. H. CARPENTER, AND R. A. GARROTT. 1987. Aerial mark-recapture estimates of confined mule deer in Pinyon-juniper woodland. J. Wildl. Manage. 51:41-46.

BROWNIE, C. 1987. Recent models for mark-recapture and mark-resighting data. Biometrics 43:1017-1019.

_____, AND D. S. ROBSON. 1983. Estimation of time-specific survival rates from tag-resighting samples: a generalization of the Jolly-Seber model. Biometrics 39:437-453.

CHAO, A. 1988. Estimating animal abundance with capture frequency data. J. Wildl. Manage. 52:295-300.

_____. 1989. Estimating population size for sparse data in capture-recapture experiments. Biometrics 45:427-438.

DARROCH, J. N., AND D. RATCLIFF. 1980. A note on capture-recapture estimation. Biometrics 36:149-153.

GERRARD, J. M., G. R. BORTOLOTTI, E. H. DZUS, P. N. GERRARD, AND D. W. A. WHITFIELD. 1990. Boat census of bald eagles during the breeding season. Wilson Bull. 120:720-726.

JOHNSON, N. L., AND S. KOTZ. 1969. Distributions in statistics: discrete distributions. Houghton Mifflin Co., Boston. 328pp.

LEWONTIN, R. C., AND T. PROUT. 1956. Estimation of the number of different classes in a population. Biometrics 12:211-223.

MANLY, B. F. J. 1984. Obtaining confidence limits on parameters of the Jolly-Seber model for capture-recapture. Biometrics 40:749-758.

MINTA, S., AND M. MANGEL. 1989. A simple population estimate based on simulation for capture-recapture and capture-resight data. Ecology 70:1738-1751.

MORGAN, B. J. T., AND S. N. FREEMAN. 1989. A model with first-year variation for ring-recovery data. Biometrics 45:1087-1101.

OTIS, D. L., K. P. BURNHAM, G. C. WHITE, AND D. R. ANDERSON. 1978. Statistical inference from capture data on closed animal populations. Wildl. Monogr. 62:1-135.

SCHWARZ, C. J., K. P. BURNHAM, AND A. N. ARNASON. 1988. Post-release stratification in band-recovery models. Biometrics 44:765-785.

SEBER, G. A. F. 1970. The effects of trap response on tag recapture estimates. Biometrics 26:13-22.

_____. 1982. The estimation of animal abundance and related parameters. Second ed. Griffin, London. 654pp.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry. Second ed. W. H. Freeman and Co., San Fransisco, Calif. 859pp.

SPROTT, D. A. 1981. Maximum likelihood applied to a capture-recapture model. Biometrics 37:371-375.

APPENDIX

Derivation of the Likelihood and its Properties

The likelihood function, L(*N,M*) is the expression for the sampling probability of the $\{f_i\}$ given *n* in terms of the unknown parameters *N* and *M*. This can be expressed as the product of the probability of obtaining m marked sightings out of n and the conditional probability of the $\{f_i\}$ given *m*; this second probability depends on *M* alone:

$$L(N, M) = P\{m \mid n; N, M\} \cdot P\{f_1, f_2, \dots \mid m; M\}$$
$$= P_1 (N, M) \cdot P_2 (M) .$$

Because the probability of sighting a marked animal is *M/N* at each sighting, the probability of having m marked sightings among the n sightings (given our assumption that animals are sighted independently and with equal and constant probability) is given by the binomial distribution:

$$P_1(N, M) = \binom{n}{m}\left(\frac{M}{N}\right)^m \left(1 - \frac{M}{N}\right)^{n-m} .$$

The second term $P_2(M)$ is an instance of a class of complicated probability problems called occupancy problems. Fortunately, $P_2(M)$ is the likelihood developed by Lewontin and Prout (1956) (their eq 5; for comparison note that their symbols *n, k, N,* and *m* stand for what we define as *M, m′,* m, and f_0, respectively). Grouping all terms not involving M into a coefficient *K* gives:

$$P_2(M) = K \cdot M ! / [M^m \cdot (M - m′) !] .$$

From the form of L(*N, M*) we can see that all the information on *N* is contained in the first term, P_1, and that the individual f_i are not needed to form the likelihood but only the 2 statistics *m* and *m′* based on them; that is, *m* and *m′* are sufficient for inference about *N* and *M*, given *n*. We can also see that if *M* is known, then the likelihood of *N* given *n* and *M* is just P_1 and that *m* is sufficient for *N*. In this case, the maximum likelihood estimate for N is defined by the Petersen ratio (our eq 1 with *N* replaced by $\hat{N}$).

Equations 2 and 3, which define the unconstrained maximum likelihood estimates for N and M, were derived by equating to zero the partial derivatives of l (*N, M*) = log L(*N,M*) with respect to *N* and *M*. Lewontin and Prout (1956) give a justification for this procedure (our eq 3 also defines the maximum likelihood estimate for *M* in Lewontin and Prout (1956), their eq 6) despite the fact that *M* is an integer. The variance-covariance matrix for the unconstrained estimates was derived as the inverse of the actual information matrix: the matrix of negative second partial derivatives of l, evaluated at (*N, M*) = ($\hat{N}$, $\hat{M}$) and the observed (actual) values of the data (*n, m, m'*). Likelihood theory provides no guidance on how the variance- covariance matrix of the constrained estimates might be estimated. Because the main reason for obtaining the variance estimates is to form a confidence interval, we circumvent the problem by using profile relative likelihoods which do not need a standard error for their construction and are easily adjusted to satisfy the constraints after they are constructed. The profile relative likelihood function for *M* is:

$$R_P(M) = L\left(\frac{Mn}{m}, M\right) / L\left(\hat{N}, \hat{M}\right).$$

That is, the likelihood function is made a function of *M* alone by using equation 1 and then scaled to take a maximum of 1 at $\hat{N}$, $\hat{M}$. The 95% profile confidence interval, (M_L, M_U), is the range of *M* such that $R_p(M) > 0.15$ (0.15 = antilog$_e${$-\chi^2/2$} where χ^2 is the 95th percentile of a Chi-square variable with 1 *df*) and must be found by iterative search for the 2 roots of the equation $R_P(M) = 0.15$. This substitution and univariate search method works for *M* because the contour extremes in the *M* direction for any given contour lie on the straight line defined by equation 1 (Fig. 2a and c). This equation defines the *N*-isocline of R; that is, the set of (*N, M*) points where the partial derivative of R(*N,M*) with respect to *N* is zero for fixed *M*. The profile relative likelihood, $R_p(M)$, is just the height of the R(*N,M*) function along this isocline plotted against *M*. A similar substitution and search method works to find the profile confidence interval for *N*. The *M*-isocline equation, however, is nonlinear in *M* and has no explicit solution for *M* in terms of *N*, so the substitution also involves an iterative search. Computing a confidence interval for *N* with the transformation method described in this paper is simple and, because there is no

direct constraint on N, the simple method works well and probably gives results that are close to the profile confidence interval.

Derivation of the Goodness-of-Fit Test

A goodness-of-fit test for the model should be based on the distribution of the sighting frequencies $f_1, f_2 ...$ conditional on the sufficient statistics m, m'. For methods of test construction see Schwarz et al. (1988). An exact test would be difficult to construct and apply but a test that uses the same data and is conditioned in the same way can be constructed using the zero-truncated Poisson distribution (Johnson and Kotz 1969: 104-106). The test is probably asymptotically equivalent to the exact test. It is constructed as follows:

(1) The mean observed sighting frequency, given an animal is sighted, is m/m' and the mean of a zero-truncated Poisson distribution is $\mu/(1-e^{-\mu})$, where μ is the (unknown) unconditional mean sighting rate per individual ($\mu = n/N$). The maximum likelihood estimate for μ satisfies:

$$\hat{\mu} / (1-e^{-\hat{\mu}}) = m/m'$$

and can be found by trial and error search ($\hat{\mu}$ will be less than m/m').

(2) Next, form the conditional Poisson probabilities:

$$P_i' = \hat{P}_i / (1 - \hat{P}_o) \qquad\qquad i = 1, 2, ...$$

where
$$\hat{P}_i = \hat{\mu}^i e^{-\hat{\mu}} / i! \qquad\qquad i = 0, 1, ...$$

and then form the expected frequencies:

$$\hat{f}_i = m' P_i' \qquad\qquad i = 1, 2, ...$$

Each $\hat{f}_i$ is the correctly conditioned expected sighting frequency because $\sum \hat{f}_i = \sum f_i = m'$ and $\sum i \hat{f}_i = \sum i f_i = m$.

(3) Finally, grouping classes as necessary, form the usual G or χ^2 statistic for testing the fit of the distribution (Sokal and Rohlf 1981:714-715). The statistic is judged against the χ^2 distribution with $a - 2$ degrees of freedom where $\underline{a}$ is the number of classes after grouping. There must be >2 sighting classes to form the test.

Table 1. Simulations of surveys with n sightings in a population of size *N* with *M* marked animals. Means (AV) and standard deviations (SD) are over 900 runs, and lines in bold indicate surveys that can be expected to yield acceptable estimates.

			Bad[a] because		Statistics			Estimates			Rel. prec[c].		
			$m'=m$	$m'=1$	m'	m	Inad[b].	$\hat{N}$		$\hat{S}(\hat{N})$	CIQW	CV	cover[d]
n	*N*	*M*	%	%	AV	AV	%	AV	SD	AV	%	%	%
50	50	3	37.2	11.8	2.4	4.3	73.2	30.2	13.8	23.5	264	78	87
		6	16.5	1.0	4.1	6.7	19.0	46.2	25.1	33.3	239	72	95
		12	2.4	0	7.6	12.0	0.1	59.1	36.6	34.9	164	59	94
	100	3	75.7	10.4	2.2	3.6	77.2	34.1	13.3	29.2	322	86	75
		6	56.8	4.7	3.1	4.6	35.4	48.2	22.2	40.7	328	84	87
		12	32.1	0	5.1	7.0	5.2	76.8	38.2	63.1	329	82	94
100	50	3	6.5	4.3	2.7	6.4	82.0	43.2	24.6	26.6	159	62	88
		6	0.6	0	5.2	11.6	43.6	54.7	31.4	25.0	90	46	92
		12	0	0	10.4	23.9	6.0	51.1	15.3	12.9	28	25	91
	100	3	40.0	11.5	2.4	4.3	72.9	60.5	28.3	47.8	274	79	87
		6	15.6	0.8	4.0	6.7	20.8	92.1	50.5	67.4	251	73	95
		12	2.5	0	7.8	12.3	0.7	116.7	74.0	69.5	171	60	95
200	50	3	0	0	3.0	11.9	98.6	42.1	20.0	15.2	49	36	84
		6	**0**	**0**	**5.9**	**23.8**	**98.8**	**46.7**	**12.3**	**10.3**	**23**	**22**	**88**
		12	**0**	**0**	**11.8**	**48.2**	**96.1**	**48.3**	**7.9**	**6.6**	**14**	**14**	**88**
	100	3	6.8	3.9	2.7	6.5	81.8	88.1	51.0	55.3	169	63	86
		6	**0.2**	**0**	**5.2**	**11.9**	**47.9**	**103.4**	**58.2**	**45.8**	**84**	**44**	**92**
		12	**0**	**0**	**10.4**	**23.9**	**4.3**	**102.1**	**31.6**	**26.9**	**29**	**26**	**94**

[a] Bad runs yield no estimates; each bad run was regenerated until one yielding estimates occurred.

[b] Percent of the 900 runs yielding inadmissible $\hat{M} < m'$.

[c] Relative precision as measured by CIQW = 100 AV[confidence interval quarter width] / AV[$\hat{N}$], and by coefficient of variation, CV = 100 AV[$\hat{S}(\hat{N})$] / AV[$\hat{N}$].

[d] Percent of 900 runs where a nominal 95% confidence interval covers *N*.

Table 2. Statistics, estimates, and goodness-of-fit tests for 3 data sets.

	Statistics[a]			(a) Estimates							
				M				N			
Species	n	m	m'	Est.[c]	Min.[d]	SE	95% CI	Est.[c]	Min.[d]	SE	95% CI
Eagles	107	7	3	2.59	3	0.65	2.0–5.8	39.6	45.9	17.6	18.5–110.7
Hypothetical	100	21	9	9.60	9	1.37	8.1–14.2	45.7	42.9	11.0	29.5–76.4
Badgers	107	68	13	12.4	13	0.34	12.0–13.5	19.4	20.5	1.5	16.7–22.8

(b) Goodness-of-fit[b]

	$\hat{\mu}$ [e]	χ^2	df	P
Eagles	2.026	0.03	0	--
Hypothetical	2.026	1.06	2	0.59
Badgers	5.202	4.58	4	0.33

[a] Capture frequencies $f_1, f_2,...$ were, for eagles: 0, 2, 1, 0,...; for hypothetical: 2, 3, 3, 1, 0,...;
and for badgers: 2, 1, 1, 0, 3, 2, 2, 0, 1, 0, 1, 0,...

[b] Sighting frequency class poolings used were, for eagles: $\leq 2, \geq 3$; for hypothetical: 1, 2, 3, ≥ 4);
 and for badgers: 1-2, 3-4, 5, 6, 7, ≥ 8.

[c] Unconstrained maximum likelihood estimate.

[d] Minimum constraint boundary for estimate.

[e] Estimated unconditional mean sighting rate per individual for the zero-truncated Poisson distribution
fitted to the sighting frequency data.

HISTOGRAM OF UNTRANSFORMED $\hat{N}$

HISTOGRAM OF TRANSFORMED $\hat{N}$
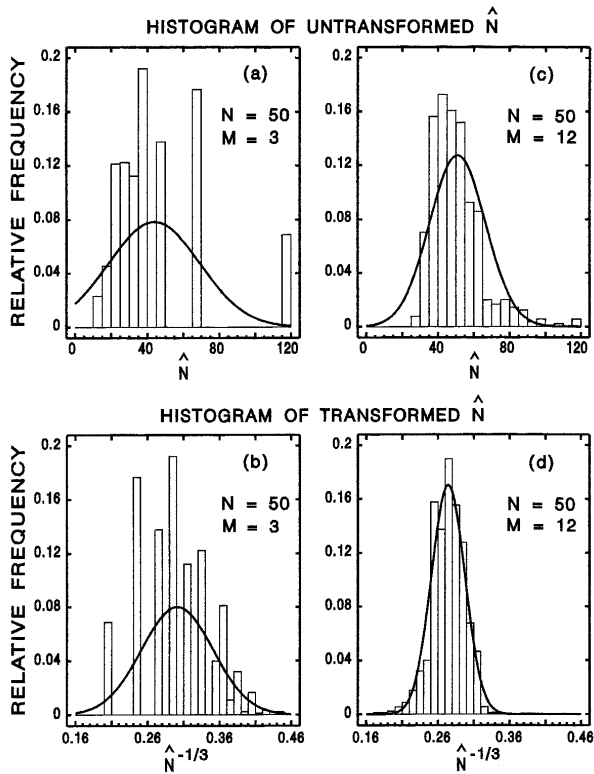
Fig. 1. Distribution of $\hat{N}$ over 900 simulations of a survey involving 100 sightings in a population of (true) size $N = 50$ and containing $M = 3$ in (a) and (b) or $M = 12$ in (c) and (d), marked birds. For small M, the discontinuities in the distribution result from the few discrete possibilities for m and m'. Use of the inverse cube root transformation in (b) and (d) evens out the discontinuities and improves the fit to a normal distribution, shown as a solid curve.

JOINT RELATIVE LIKELIHOOD CONTOUR PLOT FOR N,M

PROFILE RELATIVE LIKELIHOOD PLOT FOR M
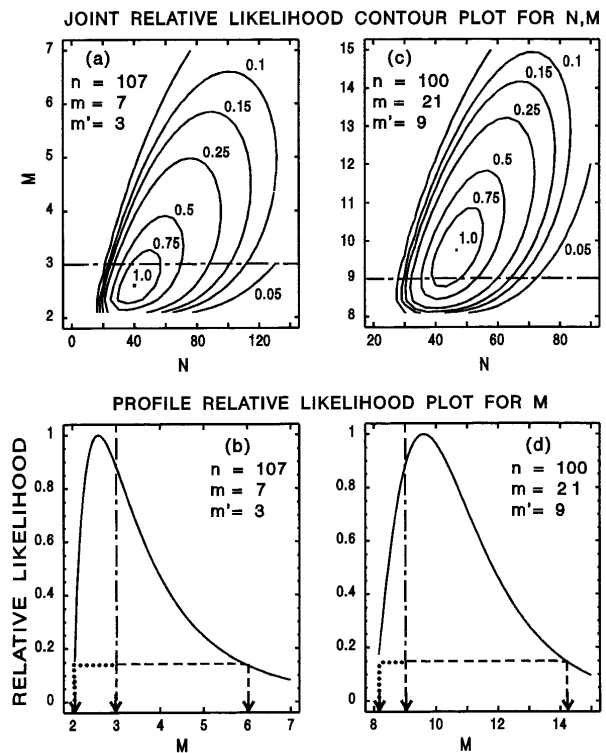
Fig. 2. Joint relative likelihoods R(N, M) (top) and profile relative likelihood $R_P(M)$ (bottom) for eagle data (at left) and hypothetical data (at right) where N is population size and M is number marked. Maximum likelihood estimates occur where R = 1; solid lines in (a) and (c) give contours of constant relative likelihood at values indicated by the numeric labels. Estimates of M below the value indicated by $-\cdot-$ are inadmissible. Dashed lines and arrows in (b) and (d) indicate the 95% profile confidence interval for M, which should properly exclude inadmissible values (dotted line).