Ecology, 59(1), 1978, pp. 190–193
© 1978 by the Ecological Society of America

A LOG-LINEAR MODEL APPROACH TO ESTIMATION OF POPULATION SIZE USING THE LINE-TRANSECT SAMPLING METHOD[1]

D. R. ANDERSON,
Utah Cooperative Wildlife Research Unit, Utah State University, Logan, Utah 84322 USA

K. P. BURNHAM,
U.S. Fish and Wildlife Service, Post Office-Federal Building, Fort Collins, Colorado 80521 USA

AND

B. R. CRAIN
Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019 USA

Abstract. The technique of estimating wildlife population size and density using the belt or line-transect sampling method has been used in many past projects, such as the estimation of density of waterfowl nesting sites in marshes, and is being used currently in such areas as the assessment of Pacific porpoise stocks in regions of tuna fishing activity. A mathematical framework for line-transect methodology has only emerged in the last 5 yr. In the present article, we extend this mathematical framework to a line-transect estimator based upon a log-linear model approach.

Key words: Density; line-transects; population estimation; population size; sampling.

INTRODUCTION

Let a region of known area A contain an unknown number, N_A, of animals. Let a rectangular strip of length L and width 2w be placed at random within the area A. The transect strip itself has area 2Lw and contains an unknown number, N, of animals. An observer walks along the center line of the transect strip (for a distance of L miles) and records n, the number of animals observed, and $Z_1, Z_2, \ldots, Z_n$, the right-angle or perpendicular distances of the observed animals from the center line of the transect. (It will be immaterial whether an animal is seen to the left or the right of the center line, so that distances are absolute and $0 \leq Z_j \leq w$ for $j = 1, 2, \ldots, n$. Animals seen beyond the boundaries of the transect strip are not recorded.)

Based on the data collected, an estimator $\hat{N}$ of N is constructed, and then an estimator $\hat{N}_A$ of N_A is given by $\hat{N}_A = (A/2Lw)\hat{N}$. If D is the density of animals in the region, a simple estimate of D is of the form $\hat{D} = \hat{N}/2Lw$.

The totality of assumptions made in this article follows exactly the treatment of Burnham and Anderson (1976). The reader may wish to refer to that paper as well as some of the other references listed on line-transect sampling. The main assumption is that animals are independently and uniformly distributed over the region (and hence throughout the transect strip). Finally, we remark that the term "animals" is being used in a very loose sense, since the observer may be counting almost anything that meets the assumptions given in Burnham and Anderson (1976), such as big game animals, plants, birds, waterfowl nests, winter-killed deer, etc.

[1] Manuscript received 4 November 1976; accepted 31 May 1977.

By way of example, Anderson and Pospahala (1970) presented data on waterfowl nests gathered on the Monte Vista National Wildlife Refuge, Colorado, during 1967–1968. Approximately A = 4,047 hectares of the refuge were sampled by the line-transect method. Transect strips 4.88 m wide were conducted for a total length of 2,574.4 km. Thus L = 2,574.4 km and w = 2.44 m.

A grand total of n = 534 waterfowl nests were observed within the transect strips. Perpendicular distances were measured and the data were grouped as follows. The interval [0,8] was partitioned into subintervals [0,1], [1,2], . . . , [7,8]. The number of class counts for these intervals (in order) were 74, 73, 79, 66, 78, 58, 52, 54 respectively. Note that the data tend to decay or drop off away from the center line of the transect strip. This example and other similar data sets lead to the following general formulation.

Under the Burnham and Anderson (1976) framework (see also Gates et al. [1968]), there is a function g(x) which quantifies the increasing difficulty of seeing animals which are farther and farther away from the center line of the transect. More precisely, g(x) is given by

$$g(x) = \Pr\,(\text{observing an animal} \mid \text{its perpendicular distance from center line is } x),\ 0 \leq x \leq x.$$

It is assumed that g(x) is monotone nonincreasing over [0,w], that g(x) is a positive, continuous funtion, and furthermore that g(0) = 1.

Burnham and Anderson (1976) show that the right-angle distance measurements $Z_1, Z_2, \ldots, Z_n$ may be regarded as independent, identically distributed random variables with (common) probability density function f(z) given by

$$f(z) = g(z)/\mu_w,\ 0 \leq z \leq w,$$

where

$$\mu_w = \int_0^w g(u)du.$$

For a randomly selected animal in the strip, the probability that the animal will be observed is $\bar{g}$, where

$$\bar{g} = 1/w \int_0^w g(x)dx = \mu_w/w.$$

Another notation for $\bar{g}$ is $P_w = \mu_w/w$. Notice also that $\bar{g} = E[g(\tilde{X})]$, where $\tilde{X}$ is distributed as uniform [0,w].

Now the following equations lead to natural estimates of N and D:

$$E(n) = NP_w = N\bar{g} = N\mu_w/w,$$

$$f(0) = g(0)/\mu_w = 1/\mu_w$$
(because g[0] = 1 by assumption),

$$N = wE(n)/\mu_w = wE(n)f(0),$$

$$D = N/2Lw = E(n)/2L\mu_w = [E(n)f(0)]/2L.$$

Thus, as given in Burnham and Anderson (1976), we have

$$\hat{N} = wn\hat{f}(0) \quad \text{and} \quad \hat{D} = n\hat{f}(0)/2L.$$

In this article, $\hat{f}(0)$ is an estimator of f(0), the probability density function (pdf) f(z) of $\tilde{Z}$ (*observed* right-angle distance) evaluated at z = 0.

It should be noted that f(z) is **not** uniform [0,w], but rather $f(z) = g(z)/\mu_w$, $0 \leq z \leq w$, because it refers only to animals that are **observed**. Because f(z) is proportional to g(z), and g(z) is nonincreasing, then f(z) is nonincreasing. This, in turn, implies that $f(0) \geq 1/w$, because 1/w is the average value of f(z) over [0,w].

The method of this paper is to develop an estimate $\hat{g}(x)$ of the function g(x), and hence obtain estimates of f(0), $\bar{g}$, and μ_w. The method used is a log-linear model approach. Gates et al. (1968) study the censusing of grouse populations and assume the parametric form $g(x) = \exp(-\lambda x)$, $\lambda > 0$. Anderson and Pospahala (1970) study waterfowl nests and present data which is well described by taking g(x) = exp[q(x)], with q(x) a quadratic polynomial. Amman and Baldwin (1960), in studying woodpeckers, find g(x) to be constant. All of these, and other data sets we have seen, can be fitted with the model to be proposed below.

The Model and Estimation

Assume that g(x) has the exact (or perhaps approximate) analytical form

$$g(x) = \exp(a + bx + cx^2), \quad 0 \leq x \leq w.$$

Because $g(0) = e^a = 1$ by assumption, we take a = 0. If c = 0 we have the exponential model. If b = 0 we have the shape of a half-normal curve. And if b = c = 0 we get $g(x) \equiv 1$ over [0,w].

The rationale for the above model is as follows. By the physical nature of the problem, it seems reasonable to conjecture that: (1) g(0) = 1; (2) g(x) is monotone nonincreasing over [0,w]; (3) g(x) is positive; and (4) g(x) has either zero or 1 inflection point. The above model accommodates all of these properties, and gives a form for g(x) that is conceptually and analytically simple and convenient. In addition, g(x) can be made either concave or convex by proper choice of a and b.

Now, as with the waterfowl nest data of Anderson and Pospahala (1970), assume the interval [0,w] has been partitioned into k subintervals $I_1, I_2, \ldots, I_k$ of equal width w/k, so that

$$I_j\ [(j-1)w/k, jw/k], \quad j = 1, 2, \ldots, k.$$

Let n_j be the number of animals seen in the transect strip whose perpendicular distance from the center line of the transect falls into I_j, j = 1, 2, . . . , k. Then $n = n_1 + n_2 + \cdots + n_k$ is the total number of animals observed in the transect strip, and $N = n + (N - n)$ is the total (unknown) number of animals in the transect strip, including both observed and unobserved animals.

The joint distribution of $(\tilde{n}_1, \tilde{n}_2, \ldots, \tilde{n}_k, N - \tilde{n})$ is multinomial with parameters N and $p_1, p_2, \ldots, p_k, p_{k+1}$ given by

$$p_j = \bar{g}_j/k, \quad j = 1, 2, \ldots, k,$$

where $\bar{g}_j$ = average value of g(x) over I_j (see below), and

$$p_{k+1} = 1 - [(\bar{g}_1/k) + (\bar{g}_2/k) + \cdots + (\bar{g}_k/k)] = 1 - \bar{g}.$$

Recall that $\bar{g} = (1/w) \int_0^w g(x)dx$ is the average value of g(x) over [0,w]. The quantity $\bar{g}_j$ is defined to be the average value of g(x) over I_j, or

$$\bar{g}_j = (\text{length of } I_j)^{-1}\left(\int_{I_j} g(x)dx\right) = (w/k)^{-1}\left(\int_{I_j} g(x)dx\right) = (k/w)\left(\int_{I_j} g(x)dx\right).$$

Then

$$(\bar{g}_1/k) + (\bar{g}_2/k) + \cdots + (\bar{g}/k) = 1/w\left\{\int_{I_1} g(x)dx + \int_{I_2} g(x)dx + \cdots + \int_{I_k} g(x)dx\right\} = 1/w\int_0^w g(x)dx = \bar{g}.$$

We may write

$$\Pr(\tilde{n}_1 = n_1, \tilde{n}_2 = n_2, \ldots, \tilde{n}_k = n_k) = N!/(n_1!n_2!\ldots n_k!N - n!)(\bar{g}_1/k)^{n_1}(\bar{g}_2/k)^{n_2}\ldots(\bar{g}_k/k)^{n_k}(1 - \bar{g})^{N-n},$$

and consequently we have n_j distributed marginally as

$Bi(N,\bar{g}_j/k)$, $j = 1, 2, \ldots, k$. Also n is $Bi(N,\bar{g})$ and $N - n$ is distributed as $Bi(N,1 - \bar{g})$. The parameters in the distribution of n_j are easily explained. Each of the N animals in the transect strip may be thought of as a separate, independent trial, with 1/k being the probability that an animal will be in I_j, and $\bar{g}_j$ being the probability that an animal is observed, given that it is in I_j. Thus the probability that an animal contributes to the count n_j is $(1/k)\bar{g}_j$ for $j = 1, 2, \ldots, k$. The probability that it goes undetected (or contributes to the count $N - n$) is $1 - \bar{g}$.

Now let $\xi_j = [(j - 1)w/k + jw/k]/2$ be the midpoint of I_j for $j = 1, 2, \ldots, k$. By assumption,

$$g(\xi_j) = \exp(b\xi_j + c\xi_j^2), \quad j = 1, 2, \ldots, k,$$

and an estimate of $g(\xi_j)$ is given by

$$\widehat{g(\xi_j)} = n_j/n_1, \quad j = 1, 2, \ldots, k,$$

where it has been implicitly assumed that enough data has been collected, or the grouping has been done in such a way, that each cell count n_j is nonzero. (This simply gives a histogram-type estimate of g(x) scaled so that $\widehat{g(0)} = 1 = g(0)$. Recall that $f(z) = g(z)/\mu_w$.)

By assuming that each cell count n_j is nonzero, we are thus assured that we may form the quantity $\log n_j$ for each value of j. As the reader will notice, this condition is needed in the steps that follow.

The model assumption can now be written as

$$\log(n_j/n_1) = b\xi_j + c\xi_j^2 + \epsilon_j, \quad j = 1, 2, \ldots, k,$$

where all ϵ_j are error terms. Because $\log(n_1/n_1) = 0$, we work with $\log(n_j/n_1)$ for $j = 2, 3, \ldots, k$. The error terms are assumed to have means of zero, but will have nonindependent covariance and correlation structure.

Let $\tilde{X}$ be a strictly positive random variable whose mean is μ. Then $\log \tilde{X}$ may be expressed as

$$\log \tilde{X} = \log \mu + [(\tilde{X} - \mu)/\mu] - [(\tilde{X} - \mu)^2/2\mu^2] + [(\tilde{X} - \mu)^3/3\mu^3] - \cdots.$$

Using this expression we can obtain large sample (or asymptotic) means, variances, and covariances of the random variables $\log(n_j/n_1)$, $j = 2, 3, \ldots, k$. The relevant results are listed below, where the approximation $\bar{g}_1 \doteq 1$ has been used appropriately.

$$E(n_j) = N\bar{g}_j/k, \quad 1 \leq j \leq k,$$

$$Var(n_j) = N\bar{g}_j/k[1 - (\bar{g}_j/k)], \quad 1 \leq j \leq k,$$

$$Cov(n_i,n_j) = -N\bar{g}_i\bar{g}_j/k^2, \quad 1 \leq i \neq j \leq k,$$

$$E(\log n_j) \doteq \log(N\bar{g}_j/k), \quad 1 \leq j \leq k,$$

$$Var(\log n_j) \doteq (k/N\bar{g}_j) - (1/N), \quad 1 \leq j \leq k,$$

$$Cov(\log n_i,\log n_j) \doteq -1/N, \quad 1 \leq i \neq j \leq k,$$

$$E[\log(n_j/n_1)] \doteq \log(\bar{g}_j/\bar{g}_1) \doteq \log \bar{g}_j, \quad 2 \leq j \leq k,$$

$$Var[\log(n_j/n_1)] \doteq k/N[(1/\bar{g}_j) + (1/\bar{g}_1)] \doteq k/N[(1/\bar{g}_j) + 1], \quad 2 \leq j \leq k,$$

$$Cov[\log(n_i/n_1),\log(n_j/n_1)] \doteq k/N\bar{g}_1 \doteq k/N, \quad 2 \leq i \neq j \leq k.$$

If we make the identifications $Y_j = \log(n_j/n_1)$ for $2 \leq j \leq k$, $Y = (Y_2, Y_3, \ldots, Y_k)'$, $\beta = (b, c)'$, $\epsilon = (\epsilon_2, \epsilon_3, \ldots, \epsilon_k)'$, and define X to be the $(k - 1) \times 2$ matrix whose jth row is (ξ_{j+1},ξ_{j+1}^2), then we have the linear model

$$Y = X\beta + \epsilon,$$

where ϵ has $(k - 1)$ by $(k - 1)$ covariance matrix Σ. Written out in full, the design matrix X is given by

$$X = \begin{bmatrix} \xi_2 & \xi_2^2 \\ \xi_3 & \xi_3^2 \\ \vdots & \vdots \\ \xi_k & \xi_k^2 \end{bmatrix},$$

and the covariance matrix Σ is

$$\Sigma \doteq \begin{bmatrix} k/N\left(\frac{1}{\bar{g}_2} + 1\right) & k/N & \cdots & k/N \\ k/N & k/N\left(\frac{1}{\bar{g}_3} + 1\right) & \cdots & k/N \\ \vdots & & \ddots & \vdots \\ k/N & k/N & \cdots & k/N\left(\frac{1}{\bar{g}_k} + 1\right) \end{bmatrix}$$

It is evident that $(N/k)\Sigma \doteq \Lambda + 11'$, where

$$\Lambda = \begin{bmatrix} 1/\bar{g}_2 & 0 & \cdots & 0 \\ 0 & 1/\bar{g}_3 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\bar{g}_k \end{bmatrix}$$

and $1 = (1, 1, \ldots, 1)' \in R^{k-1}$. Using a result from Rao (1965), we have

$$(\Lambda + 11')^{-1} = \Lambda^{-1} - (1 + 1'\Lambda^{-1}1)^{-1}\Lambda^{-1}11'\Lambda^{-1},$$

and after some matrix operations we obtain an expression for the inverse of the covariance matrix Σ as

$$\left(\frac{\bar{g}}{N}\right)\Sigma^{-1} \doteq \begin{bmatrix} \frac{\bar{g}_2}{k}\left[\bar{g} - \frac{\bar{g}_2}{k}\right] & -\frac{\bar{g}_2\bar{g}_3}{k^2} & \cdots & -\frac{\bar{g}_2\bar{g}_k}{k^2} \\ -\frac{\bar{g}_3\bar{g}_2}{k^2} & \frac{\bar{g}_3}{k}\left[\bar{g} - \frac{\bar{g}_3}{k}\right] & \cdots & -\frac{\bar{g}_3\bar{g}_k}{k^2} \\ \vdots & & \ddots & \vdots \\ -\frac{\bar{g}_k\bar{g}_2}{k^2} & -\frac{\bar{g}_k\bar{g}_3}{k^2} & \cdots & \frac{\bar{g}_k}{k}\left[\bar{g} - \frac{\bar{g}_k}{k}\right] \end{bmatrix}.$$

(In the above derivation we have used $\bar{g}_1 \doteq 1$ and $\bar{g} = (\bar{g}_1 + \bar{g}_2 + \cdots + \bar{g}_k)/k$ in the combined form $\bar{g}_2 + \bar{g}_3 + \cdots + \bar{g}_k \doteq k\bar{g} - 1$.)

If Σ were known, standard linear models theory (e.g., Graybill [1976]) would lead to the estimator $\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$. The covariance matrix Σ is unknown, but can be estimated. One could take $\hat{\bar{g}}_i =$

n_1/n_1, $\hat{\bar{g}}_2 = n_2/n_1, \ldots, \hat{\bar{g}}_k = n_k/n_1$, and for $\hat{\bar{g}}$, use the fact that $k\bar{g} \doteq 1 + \bar{g}_2 + \cdots + \bar{g}_k$, so that $k\hat{\bar{g}} = 1 + \hat{\bar{g}}_2 + \cdots + \hat{\bar{g}}_k = 1 + (n_2 + \cdots + n_k)/n_1 = 1 + (n - n_1)/n_1 = n/n_1$, and so $\hat{\bar{g}} = n/kn_1$.

Thus one has the estimator $\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}Y$. It should be mentioned that $\xi_2 < \xi_3 < \cdots < \xi_k$ implies X has rank equal to 2. Actually $\xi_2 < \xi_3$ forces the first two rows of X to be linearly independent. Thus $X'\hat{\Sigma}^{-1}X$ is 2×2 and of rank 2, and hence is invertible.

Once $\hat{\beta} = (\hat{b}, \hat{c})'$ has been obtained, our model then yields an estimate of g(x) as

$$\widehat{g(x)} = \exp(\hat{b}x + \hat{c}x^2), \quad 0 \leq x \leq w.$$

Using this, we can get an estimate of the parameter $\mu_w = \int_0^w g(x)dx$ by taking

$$\hat{\mu}_w = \int_0^w \exp(\hat{b}x + \hat{c}x^2)dx.$$

After a few manipulations, we have

$$\hat{\mu}_w = \sqrt{-(\pi/\hat{c})}\exp(-\hat{b}^2/4\hat{c}) \{\Phi[\sqrt{-2\hat{c}}(w + (\hat{b}/2\hat{c}))] - (-\Phi\hat{b}/\sqrt{-2\hat{c}})\},$$

provided that $\hat{c} < 0$. Here $\Phi(\cdot)$ represents the standard normal cumulative distribution function. Finally, we now have the estimates

$$\hat{N} = nw\hat{f}(0) = nw/\hat{\mu}_w$$

and

$$\hat{D} = \hat{N}/2Lw = n\hat{f}(0)/2L = n/2L\hat{\mu}_w.$$

Extensions

There are several factors which contribute to the variation of the estimates $\hat{N}$ and $\hat{D}$, and one of these is the condition that the covariance matrix Σ is unknown and hence must be estimated. The unknown quantities in Σ were $\bar{g}_2, \bar{g}_3, \ldots, \bar{g}_k$. We now suggest several ways in which the weighting matrix Σ may be "smoothed," an operation which should reduce the variation of the estimates $\hat{N}$ and $\hat{D}$.

One possibility is to fit a quadratic polynomial to the original cell counts, that is, use ordinary least squares and find $\gamma_0, \gamma_1, \gamma_2$ to minimize

$$S = \sum_{j=1}^{k} [n_j - (\gamma_0 + \gamma_1\xi_j + \gamma_2\xi_j^2)]^2.$$

Once $\gamma_0, \gamma_1, \gamma_2$ have been found, we "normalize" the quadratic $\gamma_0 + \gamma_1 x + \gamma_2 x^2$ and take as our first-stage estimate of g(x) the quadratic $1 + (\gamma_1/\gamma_0)x + (\gamma_2/\gamma_0)x^2$, which clearly takes the value 1 as $x = 0$.

Then we can form (recall that ξ_j = midpoint of I_j)

$$\hat{\bar{g}}_2 = 1 + \xi_2(\gamma_1/\gamma_0) + \xi_2^2(\gamma_2/\gamma_0),$$
$$\vdots$$
$$\hat{\bar{g}}_k = 1 + \xi_k(\gamma_1/\gamma_0) + \xi_k^2(\gamma_2/\gamma_0).$$

These estimates of $\bar{g}_2, \bar{g}_3, \ldots, \bar{g}_k$ then go into Σ to give $\hat{\Sigma}$, and finally $\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}Y$ as before. The advantage here should be that $\hat{\Sigma}$ becomes less variable.

A second method would be to smooth the cell counts themselves before performing any analysis or estimation. Let $m_1, m_2, \ldots, m_k$ be "smoothed" cell counts, and let $n_1, n_2, \ldots, n_k$ be the unsmoothed counts. A typical smoothing operation would be to define

$$m_1 = \tfrac{1}{2}n_1 + \tfrac{1}{2}n_2,$$
$$m_2 = \tfrac{1}{3}n_1 + \tfrac{1}{3}n_2 + \tfrac{1}{3}n_3,$$
$$m_3 = \tfrac{1}{3}n_2 + \tfrac{1}{3}n_3 + \tfrac{1}{3}n_4,$$
$$\vdots$$
$$m_{k-1} = \tfrac{1}{3}n_{k-2} + \tfrac{1}{3}n_{k-1} + \tfrac{1}{3}n_k,$$
$$m_k = \tfrac{1}{2}n_{k-1} + \tfrac{1}{2}n_k.$$

One could then regard $m_1, m_2, \ldots, m_k$ as the data, or cell frequencies and base the estimation of β on these smoothed counts. In general, smoothing results in reduced variation but increased bias.

Discussion

It may be argued that $\hat{N}$ and $\hat{D}$ are approximately unbiased for large samples. Also, it is conceivable that approximate variance expressions may be found for the estimators $\hat{N}$ and $\hat{D}$ using the delta method or Taylor's Series method. However, the more fundamental method advocated by Burnham and Anderson (1976), Overton (1969), and Eberhardt (1968), which uses replication of transect lines would seem to be more practical and should give reasonable results. For an exact treatment of this method, see Burnham and Anderson (1976).

Literature Cited

Amman, G. D., and P. H. Baldwin. 1960. A comparison of methods for censusing woodpeckers in spruce-fir forests of Colorado. Ecology **41**:699–706.

Anderson, D. R., and R. S. Pospahala. 1970. Correction of bias in belt transect studies of immotile objects. J. Wildl. Manage. **34**:141–146.

Burnham, K. P., and D. R. Anderson. 1976. Mathematical models for nonparametric inferences from line transect data. Biometrics **32**:325–336.

Eberhardt, L. L. 1968. A preliminary appraisal of line transects. J. Wildl. Manage. **32**:82–88.

Gates, C. E., W. H. Marshall, and D. P. Olson. 1968. Line transect methods of estimating grouse population densities. Biometrics **24**:135–145.

Graybill, F. A. 1976. Theory and application of the linear model. Wadsworth Publ. Co., Inc., Belmont, California. 704 p.

Overton, W. S. 1969. Estimating the number of animals in wildlife populations. Pages 403–455 *in* R. H. Giles, Jr., editor. Wildlife management techniques. The Wildlife Society, Washington, D.C.

Rao, C. R. 1965. Linear statistical inference and its applications. John Wiley and Sons, Inc., New York. 522 p.