Mathematics and Statistics 5(1): 1-4, 2017
DOI: 10.13189/ms.2017.050101 http://www.hrpub.org

A Weighted Exponential Model for Grouped Line Transect Data

Fahid Al Eibood[1], Omar Eidous[2,*]

[1]Department of Statistics, Faculty of Science, King Abdulaziz University, Saudi Arabia
[2]Department of Statistics, Faculty of Science, Yarmouk University, Jordan

Abstract This paper considers a parametric model for grouped data collected via line transect technique. The weighted exponential model is studied and investigated when the data are assumed to be grouped in the intervals. The maximum likelihood method is adopted for purpose of estimation. The resultant estimator of the population abundance is compared with the corresponding estimator that developed for ungrouped data by using the Laake stakes real data.

Keywords Weighted Exponential, Line Transect Sampling, Shoulder Condition, Grouped Data, Maximum Likelihood Estimator

1. Introduction

Line transect sampling method is a distance method used to estimate a population abundance (density). By adopting this method, an investigator attempts to estimate the population density D by walking a distance L following a path through a target region of area A. The investigator counts the number of detected objects and for each one he records the observed value of the perpendicular distance X.The distance between detected object and transect line is referred to the perpendicular distance. The number of observed objects depends on the nature of a detection function $g(x)$, which is defined to be the conditional probability of observing an object given its perpendicular distance is x. Also, there is a probability density function $f(x)$ of x given that the object with perpendicular distance x has been detected. The relationship between the two functions $g(x)$ and $f(x)$is $f(x) = g(x)/\int_0^\infty g(x)dx$. If the probability of detecting an object at zero distance is one then $f(0) = 1/\int_0^\infty g(x)dx$. The general estimator of $D = nf(0)/2L$ is $\hat{D} = n\hat{f}(0)/2L$(Burnham and Anderson, 1976), where $\hat{f}(0)$ is an appropriate sample estimator of $f(0)$ computed based on the sample of perpendicular distances $X_1, X_2, \dots, X_n$, which are assumed to be independent and identically distributed.

There are some parametric models which were proposed to estimate $f(x)$ (See for example, Gates *et al*., 1968 and Ababned and Eidous, 2012). On the other hand, the nonparametric methods to estimate $f(x)$ have been more attention in the last two decates (See Chen, 1996; Eidous 2005a, 2005b, 2006, 2009, 2011a, 2011b, 2012, 2014, 2015 and Eidous and Al-Shakhatreh, 2011). However, Nonparametric methods can be used when the perpendicular distances are ungrouped, none of these nonparametric methods can be applied when the data are grouped.

2. Grouped Data and Weighted Exponential Model

Distance data can be recorded accurately or grouped. The raw data might be incomplete because errors occurred during the data collection process. Rounding errors in measurements often cause the data to be grouped to some degree, but they must be analyzed as if they had been recorded accurately, or grouped further, in attempt to reduce the effects of rounding on bias. Distances are often assigned to predetermine distance intervals, and must be then analyzed using methods developed for a analysis of grouped data**.** Burnham *et at*. (1980) gave two reasons for considering analysis of grouped data. The first one is that the grouping is sometimes necessary due to the problems of data recording or inaccurate field measurements and the second reason is that data are sometimes recorded in groups.

Suppose that n perpendicular distances are taken in the field only by intervals, or that, for some reasons, the perpendicular distances must be analyzed as grouped data. Let the n perpendicular distances are classified according to whether they fall into either one of the k intervals (groups) and let the number of observations falling in the ith interval is n_i and $n = \sum_{i=1}^{k} n_i$. In addition, if the probability that distance x_i falls into ith interval is

P_i, $i = 1,2,\dots,k$, then the distribution of $n_1, n_2, \dots, n_k$ is a multinomial distribution with probabilities $P_1, P_2, \dots, P_k$. The basic feature of a grouped data is that no individual values of the observations are recorded and the set of group frequencies n_i, ($i = 1,2,\dots,k$) contains all all the information available in the perpendicular distances sample. For such a scheme it is natural to use the unbiased relative frequency estimator $\hat{P}_i = n_i/n$ of $P_i, i = 1,2,\dots,k$.

The theory of analysis of grouped line transect data using maximum likelihood estimation was presented by Burnhum *et al*. (1980). Let $f(x)$ be the probability density function model for the ungrouped perpendicular distances, then an estimator of $f(0)$ can validly be based on the model using grouped data as in the case of ungrouped perpendicular data. If the perpendicular distances be grouped into k intervals by boundaries $c_0 < c_1 < \cdots < c_k$ then the interval (c_{i-1}, c_i) with $c_0 = 0$ and $c_k = w$ is the *i*th interval or the *i*th group, where w may be chosen to be the the largest perpendicular distance.

Because each detection object will fall into one of the k mutually exclusive intervals, the probabilities $P_1, P_2, \dots, P_k$ sum to one (i.e. $\sum_{i=1}^{k} P_i = 1$). Under the multinomial probability distribution of $n_1, n_2, \dots, n_k$, the probability of the observed frequency counts is given by

$$\mathrm{P}(n_1, n_2, \dots, n_k) = \frac{n!}{n_1!\, n_2! \dots n_k!} P_1^{n_1} P_2^{n_2} \dots P_k^{n_k}.$$

Let $g(x;\theta)$ be the detection function of the perpendicular distance X, where $g(x;\theta)$ is the weighted exponential function given by

$$g(x;\theta) = e^{-\theta x}\left(2 - e^{-\theta x}\right), x \geq 0, \theta > 0.$$

The corresponding weighted exponential probability density function of X is

$$f(x;\theta) = \frac{2\theta}{3} e^{-\theta x}\left(2 - e^{-\theta x}\right), x \geq 0, \theta > 0.$$

Ababneh and Eidous (2012) proposed and investigated the weighted exponential model to fit the ungrouped line transect data. They showed that the model has shoulder at the origin and it is monotonically decreasing with perpendicular distance x. In addition, they pointed out the performances of this model is better than the performances of the well known models ; negative exponential and half-normal models for many reasonable practical cases.

The cumulative distribution function $F(x;\theta)$ corresponding $f(x;\theta)$ is

$$\begin{aligned} F(x;\theta) &= \int_0^x f(u;\theta)du \\ &= \frac{2\theta}{3}\int_0^x \left(2e^{-\theta u} - e^{-2\theta u}\right)du \\ &= \frac{1}{3}\left(3 + e^{-2\theta x} - 4e^{-\theta x}\right) \end{aligned}$$

To link the model $f(x;\theta)$to the grouped data properly requires that $P_i's$ $(i = 1,2,\dots,k)$ are formally expressible as functions of θ. Each cell probability P_i has an expression of the form,

$$\begin{aligned} P_i &= \int_{c_{i-1}}^{c_i} f(x;\theta)dx = F(c_i;\theta) - F(c_{i-1};\theta) \\ &= \frac{1}{3}\left(e^{-2\theta c_i} - e^{-2\theta c_{i-1}} + 4e^{-\theta c_{i-1}} - 4e^{-\theta c_i}\right). \end{aligned}$$

3. The Maximum Likelihood Estimator

For a given set data, we define a log likelihood function as

$$\begin{aligned} lnL(\theta) &= ln\, P(n_1, n_2, \dots, n_k) \\ &= ln\left(\frac{n!}{n_1! \dots n_k!}\right) + \sum_{i=1}^{k} n_i lnP_i \end{aligned}$$

where $lnP_i = ln(1/3) + ln\left(e^{-2\theta c_i} - e^{-2\theta c_{i-1}} + 4e^{-\theta c_{i-1}} - 4e^{-\theta c_i}\right)$. Therefore the log likelihood function, $lnL(\theta)$ is

$$\begin{aligned} lnL(\theta) = d(n) + \sum_{i=1}^{k} n_i \Big(&ln(1/3) \\ &+ ln\left(e^{-2\theta c_i} - e^{-2\theta c_{i-1}} + 4e^{-\theta c_{i-1}}\right. \\ &\left. - 4e^{-\theta c_i}\right)\Big) \end{aligned}$$

$d(n) = ln\left(\frac{n!}{n_1! \dots n_k!}\right)$. The ML estimator of θ is the solution of the following equation

$$\frac{dlnL(\theta)}{d\theta} = \sum_{i=1}^{k} n_i \frac{dlnP_i}{d\theta} = 0,$$

which gives

$$\sum_{i=1}^{k} n_i \left(\frac{c_i e^{-\theta c_i}\left(e^{-\theta c_i} - 2\right) + c_{i-1} e^{-\theta c_{i-1}}\left(2 - e^{-\theta c_{i-1}}\right)}{e^{-2\theta c_i} - e^{-2\theta c_{i-1}} + 4e^{-\theta c_{i-1}} - 4e^{-\theta c_i}}\right) = 0.$$

An iterative numerical method such as Newton Raphson method is needed to solve the above equation. The numerical solution of the last equation gives the maximum likelihood estimator of θ ($\hat{\theta}$say). Therefore, the estimators of $f(0)$ and population abundance D based on the weighted exponential model when the data are deal as grouped are given by

$$\hat{f}_G(0) = \frac{2\hat{\theta}}{3},$$

and

$$\hat{D}_G = \frac{n\hat{f}_G(0)}{2L}.$$

4. Numerical Example

The Laake stakes data set which is given in Burnham *et al*. (1980) is re-analyzed here. For these data, the true

probability density function $f(x)$ is unknown. However, the true population density is known and equals $D = 0.00375$ stakes/m^2. Because $D = nf(0)/2L$ then the true value of $f(0)$ is 0.110294. The number of detected stakes was $n = 68$ with line transect length $L = 1000\,m$. The observed values of the 68 perpendicular distances were given in Burnham *et al*. (1980, Page: 62).

For sake of comparison, we computed the value of weighted exponential estimator $\hat{f}(0)$ (Ababneh and Eidous, 2012) for the original data (ungrouped perpendicular distances) which gives $\hat{f}(0) = 0.12647$ and the corresponding estimator of D is 0.0043 stakes/m^2. To apply our method we need to divide the data into a number of intervals (groups). To investigate the effect of the number of chosen intervals and their boundaries on the value of the proposed estimator we compute the value of $\hat{f}_G(0)$ for $k = 6, 8$ and 10 intervals. These intervals and the corresponding frequencies are listed in Table 1. All our below results were computed by using Mathematica 7. For Table (1a), the maximum likelihood estimator of θ can be obtained by solving the equation

$$\sum_{i=1}^{6} n_i \left(\frac{c_i e^{-\theta c_i}\left(e^{-\theta c_i}-2\right)+c_{i-1}e^{-\theta c_{i-1}}\left(2-e^{-\theta c_{i-1}}\right)}{e^{-2\theta c_i}-e^{-2\theta c_{i-1}}+4e^{-\theta c_{i-1}-4e^{-\theta c_i}}} \right) = 0$$

numerically with respect to θ, where $n_1 = 23$, $n_2 = 19, n_3 = 14$, $n_4 = 5$, $n_5 = 3$, $n_6 = 4$. Also $c_0 = 0$, $c_1 = 3.195$, $c_2 = 6.295$, $c_3 = 9.395$, $c_4 = 12.495$, $c_5 = 15.595$, $c_6 = 31.310$. Therefore, the maximum likelihood estimator of θ is $\hat{\theta} = 0.203346$. The corresponding maximum likelihood estimators of $f(0)$ and D are $\hat{f}_G(0) = 0.135564$ and $\hat{D}_G = 0.004609\ stakes/m^2$ (or 46.09 stakes/hectare) respectively. Similarly, for Table (1b) the maximum likelihood estimators of θ, $f(0)$ and D are $\hat{\theta} = 0.189965, \hat{f}_G(0) = 0.126643\ and\ \hat{D}_G = 0.004306\ stakes/m^2$ (or 43.06 stakes/hectare) respectively. Also, for Table (1c), $\hat{\theta} = 0.191562$ $\hat{f}_G(0) = 0.127708$ and $\hat{D}_G = 0.004342\ stakes/m^2$(or 43.42 stakes/hectare). The values of the proposed estimator seem to be close to each other for different numbers of the intervals. In addition, the proposed estimator for grouped data is also very close to the estimator developed for ungrouped data. The same data set was analyzed by Burnham *et al*. (1980) using the Fourier series method and they found $\hat{f}(0) = 0.12\,4803$ and $\hat{D} = 0.003903$ or $\hat{D} = 39.03$ stakes/hectare.

Acknowledgements

This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. The authors, therefore, acknowledge with thanks DSR technical and financial support.

Table 1. The interval boundaries of perpendicular distances of the stakes data and the corresponding frequencies (number of observations) (a) 6 intervals (b) 8 intervals and (c) 10 intervals

Number of intervals	Boundaries of intervals	Number of observations
1	0.000 -3.195	23
2	3.195-6.295	19
3	6.295-9.395	14
4	9.395-12.495	5
5	12.495-15.595	3
6	15.595-31.310	4

(a)

Number of intervals	Boundaries of intervals	Number of observations
1	0.000 – 2.501	16
2	2.501 – 5.001	22
3	5.001 – 7.501	9
4	7.501 – 10.001	9
5	10.001 – 12.501	5
6	12.501 – 15.001	2
7	15.001 – 17.501	1
8	17.501 – 31.310	4

(b)

Number of intervals	Boundaries of intervals	Number of observations
1	0.000 – 1.895	14
2	1.895 – 3.785	14
3	3.785 – 5.675	11
4	5.675 – 7.565	9
5	7.565 – 9.445	8
6	9.445 - 11.345	3
7	11.345 – 13.235	2
8	13.235 – 15.125	2
9	15.125 – 17.015	1
10	17.015 – 31.310	4

(c)

REFERENCES

[1] Ababneh, F. and Eidous, O. (2012). A weighted exponential detection function model for line transect data. Journal of Modern Applied Statistical Methods, 11(1), 144-151.

[2] Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L. Borchers, D.L., and Thomas, L. (2001). Introduction to distance sampling. Oxford university Press, Oxford.

[3] Burnham, K. P., Anderson, D. R. (1976). Mathematical models for nonparametric influences from line transect data. Biometrics, 32, 325-336.

[4] Burnham, K. P., Anderson, D. R. and Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. Wildlife Monograph, 72, 1-202.

[5] Chen, S. X. (1996). A kernel estimate for the density of a biological population by using line-transect sampling. Applied Statistics, 45, 135-150.

[6] Eidous, O. M. (2005a). Bias correction for histogram estimator using line transect sampling. Environmetrics, 16, 61-69.

[7] Eidous, O. M. (2005b). Frequency Histogram Model for Line Transect Data with and without the Shoulder Condition. Journal of the Korean Statistical Society, 34 (1), 49-60.

[8] Eidous, O. M. (2006). A Semi-parametric model for line transect sampling. Communications in Statistics-Theory and Methods, 35, 1211-1222.

[9] Eidous, O. M. (2009). Kernel method starting with half-normal detection function for line transect density estimation. Communications in Statistics-Theory and Methods, 38, 2366-2378.

[10] Eidous, O. M. (2011a). Additive Histogram Frequency Estimator for Wildlife Abundance Using Line Transect Data without the Shoulder Condition. Metron, LXIX (2), 119-128.

[11] Eidous, O. M. (2011b). Variable Location Kernel Method Using Line Transect Sampling. Environmetrics, 22, 431-440.

[12] Eidous, O. M. (2012). A New Kernel Estimator for Abundance Using Line Transect Sampling without the Shoulder Condition. Journal of the Korean Statistical Society, 41, 267-275.

[13] Eidous, O. M. (2014). Histogram Model with Plausible Parametric Detection Function for Line Transect Data. Communications in Statistics-Theory and Methods, 43, 1-12.

[14] Eidous, O. M. (2015). Nonparametric Estimation of $f(0)$ Applying Line Transect Data with and without the Shoulder Condition. Journal of Information & Optimization Sciences, 36 (4), 301-315.

[15] Eidous, O. M. and Alshakhatreh, M. (2011). Asymptotic Unbiased Kernel Estimator for Line Transect Sampling. Communications in Statistics-Theory and Methods, 40, 4353-4363.

[16] Gates, C.E., Marshall, W.H. and Olson, D.P. (1968). Line transect method of estimating grouse population densities. Biometrics, 24, 135-145.