Estimation of Plant Biomass from Quadrat Data Using the Lognormal Distribution

GARY C. WHITE

Highlight: This paper introduces a new method of estimating plant biomass from a series of randomly located quadrats. A mixture of the binomial and lognormal distributions provide a more realistic probability density function than the commonly used normal distribution. Point and interval estimators are derived using the method of Maximum Likelihood (ML). Example calculations illustrate the method.

This paper introduces a new method of estimating plant biomass based on a series of randomly located quadrats. For a particular taxon, a series of biomass weights, $x_1, \dots, x_n$, are measured on n quadrats. The x_i are assumed observations of the random variable X . Two characteristics of these data are:

- 1) the taxon does not occur on every quadrat, so many of the x_i values are zero; and,
- 2) the distribution of positive values is highly skewed to the right (Fig. 1).

The usual method of estimating the biomass and a confidence interval is as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$s^2 = \left[\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 / n \right] / (n-1)$$

$$95\% \text{ C.I.} = \bar{x} \pm t_{(n-1;0.05)} s / \sqrt{n}$$

where $\bar{x}$ is the estimated mean biomass,

s^2 is the estimated $\text{Var}(x_i)$, assumed constant for all

x_i ,

$s/\sqrt{n}$ is the estimated $\sqrt{\text{Var}(\bar{x})}$, and

$t_{(n-1;0.05)}$ is a 0.05 level t-statistic with $n-1$ degrees of freedom.

The author was a research associate, Utah Cooperative Wildlife Research Unit, Utah State University, Logan 84322. Cooperating agencies included U.S. Fish and Wildlife Service, Utah State University, Utah Division of Wildlife Resources, and Wildlife Management Institute. The author's present address is Los Alamos Scientific Laboratory, Group H-8, MS 522, Los Alamos, New Mexico 87545.

He wishes to thank Pamela Camp, James Albee, and Eric Larsen, Utah State Univ., for assistance in the collection of the data used in the example, and Drs. David R. Anderson, leader, Utah Coop. Wildlife Res. Unit, and John C. Malechek, Dep. of Range Science, Utah State Univ., for constructive criticism of the manuscript. Drs. Gary Tietjen and Lawrence Bruckner, LASL, verified several of the equations taken from the literature, and Dr. Thomas Boullion, Texas Tech Univ., pointed out an error in the formula for $\text{Var}(\bar{x})$. This work was funded by Contract No. 14-16-0008-1197, Office of Biological Services, U.S. Fish and Wildlife Serv. to the Utah Coop. Wildlife Res. Unit.

Manuscript received May 30, 1977.

$\bar{x}$ is an unbiased estimate of $E(X)$. However, the construction of the 95% confidence interval is dependent on X being normally distributed for finite n . For large n , the Central Limit Theorem can be used to justify the use of the normal distribution. Unfortunately, biologists seldom have sample sizes compatible with this assumption, but basically claim a normal distribution for the random variable X to construct the confidence interval. This may result in too short an interval estimate and give the estimate of biomass an illusion of preciseness. The purpose of this paper is to present a method of estimating biomass and an appropriate confidence interval based on realistic distributional assumptions about the observed x_i .

Maximum Likelihood Estimation with the Lognormal Distribution

A realistic representation of the distribution of X (biomass weights) is as follows:

$$P[X < 0] = 0,$$

$$P[X = 0] = \delta,$$

$$P[X > 0] = 1 - \delta$$

δ is then the proportion of zero values in the sample. Positive values of X are assumed to be lognormally distributed. The assumption of lognormality is more realistic than the assumption of normality because this distribution has a range from 0 to

Fig. 1. A typical frequency distribution of biomass weights measured on randomly located quadrats. The first frequency represents the quadrats with zero biomass.

infinity; hence there is no probability that negative values will be observed. Also the lognormal distribution is skewed to the right. A second justification for the application of the lognormal distribution is based on the "law of proportionate effect." This law applies to a process of change where the random variable at any step of the process is a random proportion of the previous value of the random variable (Rustagi 1964). A random variable which is the result of a continued series of such proportional effects tends to be lognormal. For the weight of wet biomass of a species observed on a quadrat, possible proportionate effects may be rainfall, mineral content of the soil, or any of the other environmental parameters affecting the species growth. A competing probability density function might be the gamma distribution. However, the gamma distribution cannot be justified with the law of proportionate effect as can the lognormal distribution. Other skewed distributions such as the negative binomial and the Poisson are not reasonable because they assume X is a discrete variable, while biomass is clearly a continuous variable. Estimation for $X > 0$ will now be considered under the lognormal distribution, $\Lambda(\mu, \sigma^2)$.

Estimation of the parameters of the lognormal distribution has been extensively studied by Finney (1941), Aitchison and Brown (1969), and Bradu and Mundlak (1970). Let $y_i = \log(x_i)$ so that y_i has the normal probability density function

$$f(y)dy = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(y_i - \mu)^2/2\sigma^2]dy.$$

Define $\alpha = E(X)$, and $\beta^2 = \text{Var}(X)$ (for $X > 0$). The minimum variance unbiased estimators of α and β^2 are respectively

$$\begin{aligned}\hat{\alpha} &= \exp(\bar{y})g_n(1/2 s_y^2), \\ \hat{\beta}^2 &= \exp(2\bar{y})[g_n(2s_y^2) - g_n(\frac{n-2}{n-1} s_y^2)]\end{aligned}$$

where

$$\begin{aligned}\bar{y} &= \sum_{i=1}^n y_i/n, \\ s_y^2 &= \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)\end{aligned}$$

and the function

$$g_n(t) = 1 + \frac{n-1}{n}t + \frac{(n-1)^2 t^2}{n^2(n+1)2!} + \frac{(n-1)^3 t^3}{n^3(n+1)(n+3)3!} + \dots$$

The function $g_n(t)$ was first derived by Finney (1941) to correct the bias of the ML estimator of α :

$$\hat{\alpha}_{MLE} = \exp(\bar{y} + s_y^2/2)$$

because

$$E[\hat{\alpha}_{MLE}] = E(X) \cdot \exp\left[-\frac{(n-1)}{n} \frac{\sigma^2}{2}\right] \cdot \left(1 - \frac{\sigma^2}{(n-1)}\right)^{-(n-1)/2}$$

This result assumes s_y^2 defined as above, and not the ML estimator of s_y^2 . The ML estimator is consistent but has positive bias for finite samples.

The series $g_n(t)$ is slow to converge but is easily programmed for solution by computer:

$$(j^{\text{th}} + 1) \text{ term} = j^{\text{th}} \text{ term} \cdot \left\{ \frac{(n-1)^2 t}{n j (n+2j-1)} \right\} \quad (j = 2, 3, \dots)$$

The function $g_n(t)$ is tabled for $-2 \leq t \leq 2$ and for increments of n up to 5000 by Bradu and Mundlak (1970).

Bradu and Mundlak (1970) give an exact expression for $\text{Var}(\hat{\alpha})$:

$$\text{Var}(\hat{\alpha}) = \exp(2\mu) [\exp(2\sigma^2/n) G_n(\frac{\sigma^2}{2} - 1/2 \frac{\sigma^2}{n}) - \exp(\sigma^2)]$$

where $G_n(t) = \exp(2t) g_n(2(n+1)t^2/n^2)$, a useful relation provided by Evans and Shaban (1974), and $\bar{y}$ and s_y^2 are estimates substituted for the unknown parameters μ and σ^2 .

The $E(X)$ is often estimated with a geometric mean. However, note the resulting estimate is biased low, because the factor $g_n(1/2 s_y^2)$ is always greater than unity for $s_y^2 > 0$. Hence the geometric mean is not an appropriate estimator, and should not be used with lognormal data.

With the above summary of parameter estimation for the lognormal distribution, estimators for quadrat data where zero values occur will now be considered. Zero values represent the case where the species did not occur on the quadrat.

Let the population be such that there is a proportion δ of zero values and the distribution of positive values in lognormal, $\Lambda(\mu, \sigma^2)$. If X still denotes the corresponding random variable then

$$\begin{aligned}P[X < 0] &= 0 \\ P[X = 0] &= \delta \\ \text{and } P[X > x] &= \delta + (1 - \delta) \Lambda(\mu, \sigma^2) \quad (x > 0).\end{aligned}$$

We then write that X is $\Delta(\delta, \mu, \sigma^2)$ (Aitchison and Brown 1969). The multiplication property of the lognormal distribution also holds for the delta distribution (Aitchison and Brown 1969), i.e., if X_1 and X_2 are Δ -variates, then $X_1 X_2$ is also a Δ -variate. Therefore the same arguments that justify the use of the lognormal distribution for positive values can be used to justify the delta distribution as a representation of both zero and positive values.

Suppose a sample of size n is taken from $\Delta(\delta, \mu, \sigma^2)$, and that n_0 zero values occur, and the remaining $n_1 = n - n_0$ positive values are $x_i (i=1, \dots, n_1)$. Let

$$\begin{aligned}\bar{y} &= \sum_{i=1}^{n_1} y_i / n_1 \quad (n_1 > 0), \\ &= 0 \quad (n_1 = 0),\end{aligned}$$

and

$$\begin{aligned}s_y^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y})^2 \quad (n_1 > 1) \\ &= 0 \quad (n_1 = 0, 1)\end{aligned}$$

The likelihood function of the sample is derived from a mixture of the binomial and lognormal distributions:

$$\binom{n}{n_0} \delta^{n_0} (1-\delta)^{n_1} (\sigma\sqrt{2\pi})^{-n_1} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n_1} (y_i - \mu)^2\right\}$$

Aitchison and Brown (1969) show that n_0/n , $\bar{y}$, and s_y^2 are jointly sufficient estimators for δ , μ , and σ^2 , and so any function of n_0/n , $\bar{y}$, and s_y^2 is a most efficient estimator of its expectation. To obtain the most efficient estimators of $\kappa = E(X)$ and $\rho^2 = \text{Var}(X)$, functions of n_0/n , $\bar{y}$, and s_y^2 must be found whose expectations are κ and ρ^2 . These estimators, $\hat{\kappa}$ and $\hat{\rho}^2$, are given by Aitchison and Brown (1969):

$$\hat{\kappa} = \frac{n_1}{n} \exp(\bar{y}) g_{n_1}(1/2 s_y^2) \quad (n_1 > 1)$$

$$= \frac{x_1}{n} \quad (n_1 = 1)$$

$$= 0 \quad (n_1 = 0)$$

and

$$\hat{\rho}^2 =$$

$$\frac{n_1}{n} \exp(2\bar{y}) \left[g_{n_1}(2s_y^2) - \frac{n_1-1}{n-1} g_{n_1}\left(\frac{n_1-2}{n_1-1} s_y^2\right) \right] \quad (n_1 > 1)$$

$$= \frac{x_1^2}{n} \quad (n_1 = 1)$$

$$= 0 \quad (n_1 = 0)$$

For large n

$$\text{Var}(\hat{\kappa}) = \frac{\alpha^2}{n} \left\{ \delta(1-\delta) + 1/2(1-\delta)^2 (2\sigma^2 + (\sigma^2)^2) \right\}$$

Note that this formula differs slightly from that given by Aitchison and Brown (1969).

A 95% confidence interval on the estimated biomass is then calculated based on the asymptotic normality of the ML estimator:

$$\hat{\kappa} \pm 1.96\sqrt{\text{Var}(\hat{\kappa})}$$

The above estimator, $\hat{\kappa}$, is the average biomass per quadrat, and will be termed the delta estimator in the remaining sections. The term "delta" was provided by Aitchison and Brown (1969), and is a poor name because of confusion with other mathematical and statistical procedures of the same name. Also no information is provided with this name that the distribution is actually a mixture of the binomial and lognormal distributions.

Example Calculations

A sample data set for alfalfa (*Medicago*) will be used to illustrate the calculations. Thirty-nine samples were taken

($n=39$) with 6 positive values ($n_1=6$, $n_0=33$). The positive values are 0.5, 1.5, 2.0, 5.0, 5.5, and 6.5. Thus $\bar{y} = 0.932$ and $s_y^2 = 0.983$. Using these values the estimate of α is 3.74, $\hat{\beta}^2 = .13.67$, and $\hat{\text{Var}}(\hat{\alpha}) = 4.46$. The estimates derived from the delta distribution are $\hat{\kappa} = 0.576$ and $\hat{\text{Var}}(\hat{\kappa}) = 0.059$. A 95% confidence interval is then [0.010–1.052].

Results and Discussion

Occasionally, the estimates from the delta distribution are much larger than those from the usual method, although both are based on exactly the same data. In these cases the frequency distribution of the positive values is usually highly skewed, with several very large values. This is particularly common when sampling shrubs. Many quadrats with zero biomass values occur, and also several quadrats with very large biomass values. An example is a series of 30 quadrats for sagebrush (*Artemisia tridentata*). Nineteen positive values occurred, with $\bar{x} \pm \text{SE}(\bar{x})$ being 806 ± 292 , while $\hat{\kappa} \pm \text{SE}(\hat{\kappa})$ was 3260 ± 3026 . The lognormal distribution tends to weight large positive values more heavily, because there is a larger probability that such extreme values will occur with this distribution over the normal distribution.

The two estimators provide very similar point and interval estimates when few zero values occur, and when the frequency distribution of the x_i 's is not highly skewed. An example of this is a series of 29 quadrats taken on miscellaneous grasses (Graminae). No zero values occurred, and $\bar{x} \pm \text{SE}(\bar{x})$ was 161 ± 9 while $\hat{\kappa} \pm \text{SE}(\hat{\kappa})$ was 162 ± 11 .

Although the usual method often provides the smallest standard errors, these are probably negatively biased due to the normality assumptions made. That is, the interval estimate constructed with this standard error will include the true biomass less often than expected because of incorrect assumptions, and provide an illusion of preciseness. If the frequency distribution of the positive values is truly lognormal, the unbiased modified maximum likelihood estimates derived from the delta distribution are known to be the minimum variance estimates.

Literature Cited

- Aitchison, J., and J. A. C. Brown. 1969. The lognormal distribution. Cambridge Univ. Press, Cambridge, Mass.
- Bradu, D., and Y. Mundlak. 1970. Estimation in lognormal linear models. J. Amer. Statist. Ass. 65:198-211.
- Evans, I. G., and S. A. Shaban. 1974. A note on estimation in lognormal models. J. Amer. Statist. Ass. 69:799-781.
- Finney, D. J. 1941. On the distribution of a variate whose logarithm is normally distributed. Supplement to the J. Royal Statist. Soc. 7(2): 155-161.
- Rustagi, J. S. 1964. Stochastic behavior of trace substances. Archives of Environ. Health 8:68:76.