Environ Ecol Stat (2008) 15:175–189
DOI 10.1007/s10651-007-0046-8

Confidence intervals for the mean of the delta-lognormal distribution

David Fletcher

Received: 16 February 2005 / Revised: 25 August 2006 / Published online: 23 October 2007

Abstract Data that are skewed and contain a relatively high proportion of zeros can often be modelled using a delta-lognormal distribution. We consider three methods of calculating a 95% confidence interval for the mean of this distribution, and use simulation to compare the methods, across a range of realistic scenarios. The best method, in terms of coverage, is that based on the profile-likelihood. This gives error rates that are within 1% (lower limit) or 3% (upper limit) of the nominal level, unless the sample size is small and the level of skewness is moderate to high. Our results will also apply to the delta-lognormal linear model, when we wish to calculate a confidence interval for the expected value of the response variable, given the value of one or more explanatory variables. We illustrate the three methods using data on red cod densities, taken from a fisheries trawl survey in New Zealand.

Keywords Lognormal · Profile likelihood · Skewness · Zero-inflated data

1 Introduction

In the life sciences it is common for data to be skewed and to contain a relatively large number of zeros. In some situations, it is appropriate to model such data using a lognormal distribution for the positive values, together with an additional probability mass at zero. This type of distribution is commonly referred to as the delta-lognormal, and was first discussed by Aitchison (1955). It has been used in various applications since then, and is well-known in fisheries research (Pennington 1983,1991;

D. Fletcher (✉)
Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand
e-mail: dfletcher@maths.otago.ac.nz

D. Fletcher
Proteus Wildlife Research Consultants, P.O. Box 5193, Dunedin, New Zealand

Smith 1988,1990; Myers and Pepin 1990; Lo et al. 1992). We consider situations in which the parameter of interest is the mean of the distribution, rather than the median. This will be the case, for example, when we are interested in the mean density of a species, as it is directly related to total abundance.

In this paper, we consider how to calculate a 95% confidence interval for the mean of the delta-lognormal distribution. We use simulation to compare three methods for calculating such an interval. Previous research on this distribution has tended to focus on the properties of different point estimators of the mean, rather than on interval estimation (Pennington 1983,1991; Myers and Pepin 1990).

Our results will also be applicable to the more general situation where we have a delta-lognormal linear model (Lo et al. 1992; Stefansson 1996), and wish to calculate a confidence interval for the expected value of the response variable, given the value of one or more explanatory variables. We consider the simpler case, with no explanatory variables, for ease of exposition.

Note that we do not focus on the robustness of the confidence interval to model-misspecification. Our aim is simply to assess the performance of different methods when a delta-lognormal distribution is appropriate.

2 Fisheries example

In order to illustrate the type of data we are concerned with, consider the following example, taken from a trawl survey carried out by the National Institute of Water & Atmospheric Research in New Zealand. The data we focus on are the density (kg/km^2) of red cod (*Pseudophycis bachus*) for a sample of 67 trawls. As mentioned above, for ease of presentation we do not consider covariates that might influence density. Likewise, we do not consider any stratification of the area, and focus simply on estimation of the mean density of red cod in the study area. There were 13 trawls (19%) with no red cod; Fig. 1 shows the distribution of the 54 positive densities, both on the original scale and after log-transformation (these densities are also shown in full in Appendix A). The Anderson-Darling goodness-of-fit statistics have P-values of 0.004 and 0.1, for the original scale and the log-scale respectively. We therefore assume a delta-lognormal distribution for the red cod densities in this area.

3 Notation and methods

Consider a non-negative random variable Y for which $\Pr(Y > 0) = \pi$. We define $X = \{\ln Y \,|\, Y > 0\}$, and assume that $X \sim N\left(\mu_X, \sigma_X^2\right)$. Y is said to have a delta-lognormal distribution. It can be readily shown (Aitchison 1955) that the expected value of Y is given by

$$\mu_Y = \pi \exp\left(\mu_X + \frac{\sigma_X^2}{2}\right) \tag{1}$$

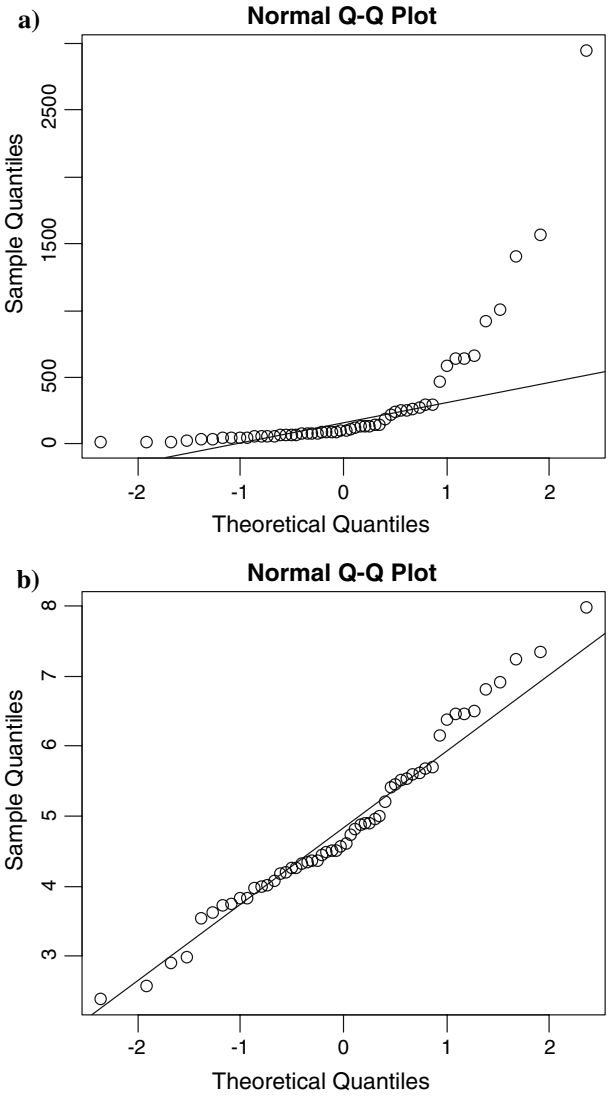

Fig. 1 Distribution of the 54 non-zero densities (kg/km^2) of red cod from a total of 67 trawls: (**a**) original scale and (**b**) after log-transformation

Suppose we have a random sample of n observations y_i $(i = 1, \ldots, n)$. Without loss of generality, we assume that the first m of these are positive, and write $x_i = \ln y_i$ $(i = 1, \ldots, m \leq n)$. Note that we restrict attention to those situations where $m > 1$, as we would be unlikely to make use of a delta-lognormal distribution if m were 0 or 1.

We compare three methods for calculating a 95% confidence interval for μ_Y. In doing so, we assume that the delta-lognormal distribution provides an appropriate model for the data, and therefore focus on parametric approaches. A traditional method for analysing skewed data that contain zeros is to use a $\ln(y+c)$ transformation, where

c is some constant (usually $c = 1$) and then proceed with a standard normal-theory analysis. It is well known that such an approach is prone to problems, in that the choice of c can affect the conclusions (Berry 1987; Mead 1988), and we do not consider its use here.

Following results presented by Finney (1941) for the lognormal distribution, Aitchison (1955) gave expressions for the minimum variance unbiased estimator (MVUE) for μ_Y. These expressions differ according to whether $m = 0$, $m = 1$ or $m > 1$. The form of the estimator for the situation we consider ($m > 1$) is

$$\hat{\mu}_Y = p \exp(\bar{x}) G_m\left(\frac{s_X^2}{2}\right) \tag{2}$$

where $p = m/n$, $\bar{x}$ and s_X^2 are the sample mean and variance of the x_i and

$$G_m(t) = 1 + \frac{(m-1)}{m}t + \sum_{i=2}^{\infty} \frac{(m-1)^{2i-1}}{m^i (m+1)(m+3)\ldots(m+2i-3)} \frac{t^i}{i!}.$$

We refer to the expression in Eq. 2 as *Aitchison's estimator*. Pennington (1983) gave expressions for the MVUE of the variance of the MVUE estimator of μ_Y, again according to whether $m = 0$, $m = 1$ or $m > 1$. The form of this estimator for the case $m > 1$ is:

$$\hat{V}(\hat{\mu}_Y) = p \exp(2\bar{x}) \left\{ p G_m^2\left(\frac{s_X^2}{2}\right) - \left(\frac{m-1}{n-1}\right) G_m\left(\frac{m-2}{m-1} s_X^2\right) \right\} \tag{3}$$

Note that the MVUE-properties of the estimators provided by Aitchison (1955) and Pennington (1983) will not carry over exactly to the expressions in Eqs. 2 and 3, but the effect of this should be negligible, as the chance of m being 0 or 1 will be small for those situations where one would use the methods presented in this paper. Assuming approximate normality for $\hat{\mu}_Y$, a 95% confidence interval for μ_Y is given by

$$\hat{\mu}_Y \pm 2\sqrt{\hat{V}(\hat{\mu}_Y)} \tag{4}$$

Clearly this will not provide the asymmetry that one would expect in a confidence interval for the mean of a skewed distribution. We have included it here because its use is implicit in the work on Aitchison's estimator (Pennington 1983,1991; Smith 1988,1990; Myers and Pepin 1990).

Land (1972) presented a comparison of a number of approximate methods for calculating a confidence interval for the mean of a lognormal distribution. The best of these was based on an idea put forward by Sir David Cox. For the delta-lognormal distribution, we modify the method as follows. We first calculate a confidence interval for $\theta = \ln \mu_Y$, and then back-transform the endpoints. From Eq. 1, we have

$$\theta = \ln \pi + \mu_X + \frac{\sigma_X^2}{2}. \tag{5}$$

A natural estimator for θ is given by the corresponding function of the statistics $p, \bar{x}$ and s_X^2, which are jointly sufficient for π, μ_X and σ_X^2. This leads to

$$\hat{\theta} = \ln p + \bar{x} + \frac{s_X^2}{2}. \tag{6}$$

Note that p is *not* the maximum likelihood estimate (MLE) for π, as we are restricting attention to those situations where $m > 1$. Thus m has a truncated binomial distribution, for which the MLE ($\hat{\pi}$) is obtained by numerical search (Finney 1949). However, the difference between p and $\hat{\pi}$ is small for those situations where one would use the methods presented in this paper. Even for m as small as 5, the relative difference between the two estimates is always less than 4% (c.f. comments above regarding Aitchison's estimator).

An estimate of the variance of $\hat{\theta}$ is given by (Appendix B):

$$\hat{V}\left(\hat{\theta}\right) \approx \frac{\left(\hat{d} - \hat{c}\right)\left(1 - \hat{c}\hat{d}\right) - m\left(1 - \hat{c}\right)^2}{m\left(1 - \hat{c}\hat{d}\right)^2} + \frac{s_X^2}{m} + \frac{s_X^4}{2\left(m + 1\right)}. \tag{7}$$

where $\hat{c} = (1 - p)^{n-1}$ and $\hat{d} = 1 + (n - 1)\, p$.

Assuming approximate normality for $\hat{\theta}$, a "back-transformed" 95% confidence interval for μ_Y is then given by

$$\exp\left\{\hat{\theta} \pm 2\sqrt{\hat{V}\left(\hat{\theta}\right)}\right\} \tag{8}$$

The final method we consider is use of a profile-likelihood interval (Venzon and Moolgavkar 1988). The profile log-likelihood is defined as

$$l_p\left(\mu_Y\right) = l\left\{\mu_Y, \tilde{\pi}\left(\mu_Y\right), \tilde{\mu}_X\left(\mu_Y\right); \mathbf{y}\right\} \tag{9}$$

where $l\left(\mu_Y, \pi, \mu_X; \mathbf{y}\right)$ is the log-likelihood, $\mathbf{y}$ is the vector of observations, and $\tilde{\pi}\left(\mu_Y\right), \tilde{\mu}_X\left(\mu_Y\right)$ are the maximum likelihood estimates of π and μ_X for *given* μ_Y. For notational convenience, we suppress the dependence of the latter two functions on μ_Y, and write

$$l_p\left(\mu_Y\right) = l\left(\mu_Y, \tilde{\pi}, \tilde{\mu}_X; \mathbf{y}\right) \tag{10}$$

The function

$$w\left(\mu_Y\right) = 2\left\{l_p\left(\hat{\mu}_Y\right) - l_p\left(\mu_Y\right)\right\}$$

has a distribution which is approximately χ_1^2, where $\hat{\mu}_Y$ is the maximum-likelihood estimate of μ_Y. This leads to 95% confidence limits being defined as the two values

of μ_Y that satisfy

$$w(\mu_Y) = 3.84, \tag{11}$$

where 3.84 is the 95th percentile of the χ^2_1 distribution. The iterative procedure used to find these values is given in Appendix C.

4 Simulations

We used simulation to assess the properties of the three methods presented above. We denote the methods as follows:

Method A: Using Aitchison's estimator (Eq. 4)
Method C: Using a modification of Cox's method for the lognormal (Eq. 8)
Method P: A profile-likelihood interval (Eq. 11)

The coverage properties of these methods are not affected by the choice of μ_X (c.f. Land 1972). This is readily seen by considering the effect of changing the value of μ_X by a specified constant, i.e. $\mu_X \to \mu_X + \delta$, say. Then we have $\mu_Y \to \exp(\delta)\,\mu_Y$ and $\bar{x} \to \bar{x} + \delta$. These transformations will not change the error rates, as can be verified by considering their effect on Eqs. 2–4, 6–8 and C.2–C.4. We therefore set $\mu_X = 0$ and considered combinations of the following factors:

1. Sample size: n = 20, 50, 100
2. Probability of a positive observation: π = 0.2, 0.5, 0.8
3. Coefficient of variation of the positive values: CV = 0.5, 1.5, 2.5.

Note that the skewness (γ) of the lognormal distribution is completely determined by its CV, via the formula $\gamma = \mathrm{CV}\left(\mathrm{CV}^2 + 3\right)$. We excluded the three combinations corresponding to $n = 20$ and $\pi = 0.2$ (with CV = 0.5, 1.5, 2.5) because for these the expected number of positive observations (m) is only 4: we would not expect any method to work well for such a small number of positive values.

We assessed the performance of each method by calculating:

1. The lower and upper error rates, i.e. the proportion of times μ_Y was below the lower limit or above the upper limit, as well as the overall error rate (lower + upper);
2. The median of both the lower and upper relative half-widths, these widths being defined as

$$\frac{\hat{\mu}_Y^L - \mu_Y}{\mu_Y} \quad \text{and} \quad \frac{\hat{\mu}_Y^U - \mu_Y}{\mu_Y}$$

respectively, where $\hat{\mu}_Y^L$ and $\hat{\mu}_Y^U$ are the lower and upper confidence limits.

For each combination, we used 10,000 simulations, giving standard errors for the lower and upper error rates of approximately 0.2%. The standard errors for the median relative half-widths were estimated using the following non-parametric bootstrap procedure. We generated 1000 bootstrap samples, each of these being a random sample

(with replacement) of size 10,000 from the original 10,000 relative half-widths. We estimated the standard error of the median using the standard deviation of the resulting 1000 bootstrap-sample medians.

5 Results

The results for the error rates are shown in Table 1. Note that where the lower error rate is given as 0.00%, it means that none of the 10,000 simulations led to the population mean being below the lower confidence limit. As would be expected, the lower error rates are generally closer to the nominal level of 2.5% than are the upper error rates. The largest departures from the nominal level are for the upper error rates, when the expected number of positive observations is small and there is a high level of skewness (corresponding to a high CV). Increasing the number of positive observations and decreasing the skewness both lead to general improvements in the error rates, as we might expect.

The worst method is clearly A: if there is a high level of skewness, the upper error rate for this method is above 8%, and the lower error rate is below 1%, even for the largest sample sizes. Method C is generally better than A, especially for the upper error rates when there is a medium to high level of skewness. Method P is the best overall. It has an error rate that is closest to the nominal level in all but two combinations for the lower limit, and all but four combinations for the upper limit. For this method, the lower error rate is always within 1% of the nominal level of 2.5%. The upper error rate is within 3% of this nominal level in all but two combinations, when π is near the boundary and the sample size is small. For Method P, the median absolute difference (across the 24 combinations) between the estimated true error rate and the corresponding nominal level is 0.4% (lower), 0.9% (upper) and 0.7% (overall): the corresponding medians for Method C are 1.2% (lower), 1.5% (upper) and 0.6% (overall).

The overall error rates show broadly the same patterns as the upper error rates, except that the performance of Method C is closer to that of Method P. In practice, we expect the lower and upper error rates to be of more direct concern than the overall error rate. For example, we would not want to conclude that a method is reliable because it gives an overall error rate that is close to the nominal level of 5%, as this may arise simply because the lower error rate is close to 0% and the upper error rate is close to 5%.

The medians of the lower and upper relative half-widths are shown in Table 2, and provide some insight into the patterns in the error rates. Taking Method P as the standard, we can see that both the lower and upper limits for Method A are too low, especially when there is a high level of skewness. For the smaller sample sizes, the upper limit for Method C tends to be too high when there is a low level of skewness and too low when there is a high level of skewness.

6 Fisheries example

In order to illustrate use of the methods presented above, we consider the red cod data. The 54 non-zero densities had a mean of 290.5 kg/km^2, and a coefficient of variation

Table 1 Estimated lower, upper and overall (lower+upper) error rates for each of three methods of calculating a 95% confidence interval for μ_Y, based on 10,000 simulations. For each combination, the expected number of positive values in the data is denoted E(m). The binomial standard errors are all less than 0.5%

n	π	E(m)	CV	Lower			Upper			Overall		
				A(%)	C(%)	P(%)	A(%)	C(%)	P(%)	A(%)	C(%)	P(%)
20	0.5	10	1.0	0.00	0.6	1.8	22.7	7.9	4.9	22.7	8.5	6.8
			0.5	0.05	1.1	2.0	17.1	6.2	5.0	17.2	7.3	6.9
			0.2	1.3	3.6	2.4	5.5	1.8	3.3	6.9	5.4	5.8
20	0.8	16	1.0	0.00	0.5	1.8	19.5	7.3	9.8	19.5	7.8	11.6
			0.5	0.04	0.9	2.1	13.1	5.2	6.0	13.2	6.1	8.1
			0.2	1.3	2.8	2.1	4.2	1.9	3.0	5.5	4.7	5.1
50	0.2	10	1.0	0.01	0.7	1.9	23.3	7.4	4.8	23.3	8.1	6.6
			0.5	0.02	1.4	1.9	16.5	5.7	4.8	16.5	7.1	6.7
			0.2	0.5	3.3	2.4	7.2	2.2	3.3	7.7	5.5	5.7
50	0.5	25	1.0	0.01	0.7	1.8	15.2	5.4	4.0	15.2	6.2	5.8
			0.5	0.06	1.4	2.2	10.3	4.3	3.4	10.4	5.6	5.6
			0.2	1.4	3.0	2.4	4.0	1.8	2.7	5.5	4.9	5.2
50	0.8	40	1.0	0.00	0.7	2.0	12.6	5.3	4.5	12.6	6.0	6.5
			0.5	0.12	1.2	2.3	9.0	4.4	3.6	9.1	5.6	5.8
			0.2	1.8	2.5	2.2	3.6	2.0	2.9	5.4	4.6	5.1
100	0.2	20	1.0	0.00	0.9	2.0	16.4	5.5	3.7	16.4	6.3	5.7
			0.5	0.03	1.2	1.8	12.4	4.6	3.9	12.4	5.8	5.8
			0.2	0.8	2.5	2.0	5.2	2.0	3.0	6.0	4.5	5.0
100	0.5	50	1.0	0.02	1.0	1.9	10.5	4.2	3.1	10.5	5.3	5.0
			0.5	0.3	1.5	2.3	7.7	3.7	3.2	8.0	5.2	5.6
			0.2	1.5	2.8	2.4	3.5	1.7	2.5	4.9	4.5	5.0
100	0.8	80	1.0	0.06	1.2	2.1	8.3	3.7	2.8	8.4	4.9	4.9
			0.5	0.4	1.5	2.4	6.1	3.1	2.6	6.5	4.6	5.0
			0.2	2.0	2.9	2.8	2.9	2.0	2.5	4.9	4.9	5.3

of 1.7. Table 3 shows 95% confidence intervals for mean density obtained using each of the three methods. Method A leads to limits that are too low, whilst those for C are slightly lower than those for P. Figure 2 shows the profile log-likelihood curve, together with the 95% confidence-level threshold corresponding to use of Method P.

7 Discussion

The aim of this paper has been to consider the coverage properties of three methods of calculating a confidence interval for the mean of a delta-lognormal distribution. Our results suggest that the profile-likelihood approach (Method P) performs reasonably

Table 2 Estimated median lower and upper relative half-widths, for each of three methods of calculating a 95% confidence interval for μ_Y, based on 10,000 simulations. For each combination, the expected number of positive values in the data is denoted E(m). The bootstrap-based standard errors are all less than 3% of the corresponding estimate

n	π	E(m)	CV	Lower			Upper		
				A	C	P	A	C	P
20	0.5	10	2.5	−1.00	−0.73	−0.68	0.65	2.40	4.43
			1.5	−0.86	−0.62	−0.60	0.63	1.48	2.12
			0.5	−0.55	−0.42	−0.45	0.52	0.73	0.65
20	0.8	16	2.5	−0.86	−0.64	−0.58	0.58	1.57	2.17
			1.5	−0.66	−0.51	−0.47	0.53	0.99	1.19
			0.5	−0.33	−0.28	−0.30	0.32	0.39	0.38
50	0.2	10	2.5	−1.03	−0.75	−0.71	0.67	2.53	4.51
			1.5	−0.91	−0.65	−0.63	0.69	1.68	2.31
			0.5	−0.65	−0.48	−0.51	0.61	0.90	0.80
50	0.5	25	2.5	−0.77	−0.57	−0.53	0.56	1.22	1.57
			1.5	−0.60	−0.46	−0.44	0.51	0.83	0.96
			0.5	−0.35	−0.29	−0.31	0.34	0.41	0.38
50	0.8	40	2.5	−0.61	−0.47	−0.43	0.49	0.86	1.02
			1.5	−0.45	−0.37	−0.34	0.38	0.55	0.62
			0.5	−0.21	−0.19	−0.20	0.20	0.23	0.22
100	0.2	20	2.5	−0.86	−0.63	−0.59	0.63	1.53	2.05
			1.5	−0.70	−0.52	−0.50	0.58	1.04	1.20
			0.5	−0.46	−0.37	−0.39	0.44	0.58	0.53
100	0.5	50	2.5	−0.57	−0.45	−0.42	0.48	0.79	0.91
			1.5	−0.43	−0.35	−0.34	0.38	0.53	0.57
			0.5	−0.25	−0.22	−0.22	0.24	0.28	0.26
100	0.8	80	2.5	−0.45	−0.36	−0.34	0.39	0.57	0.63
			1.5	−0.32	−0.27	−0.26	0.29	0.37	0.40
			0.5	−0.15	−0.14	−0.14	0.15	0.16	0.15

Table 3 95% confidence limits for mean density of red cod (kg/km^2) for each of three methods

Method	Confidence Limits	
	Lower	Upper
A	121.7	306.0
C	142.3	336.6
P	144.8	359.9

well across a broad range of scenarios, and is certainly the best of the three methods considered. Even Method P leads to error rates for the upper limit that are not ideal, but this is perhaps to be expected when we are dealing with situations that involve a relatively small amount of highly skewed data.

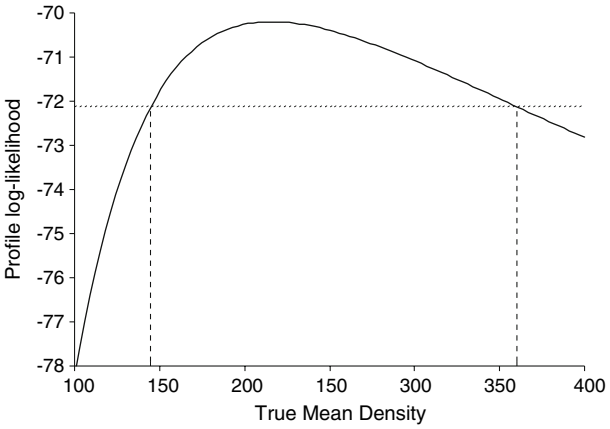

Fig. 2 Profile log-likelihood curve for true mean density of red cod (kg/km^2), together with the 95% confidence-level threshold

The approach based on Cox's method for the lognormal distribution (Method C) is the next best, although it can lead to a lower limit which is too high when there is little skewness, especially for smaller sample sizes. The use of Aitchison's estimator (Method A) is not to be recommended: it will tend to have far too low an upper limit, especially for smaller sample sizes and a high level of skewness.

Our focus has been on finding a reasonably reliable method that does not involve use of a computer-intensive procedure, such as the bootstrap or Bayes' rule. In order to check on the possible improvement in performance one might achieve using the bootstrap, we carried out a further set of simulations. These involved using a parametric studentised bootstrap (Davison and Hinkley 1997) in conjunction with Methods A and C. This led to a general improvement in their performance, but both methods were still inferior to Method P.

Throughout this paper, we have focussed on the performance of the different methods assuming that a delta-lognormal distribution is appropriate. It is possible that in some situations the positive observations will have a distribution that is more closely modelled by an alternative, such as a gamma or Weibull distribution (Dennis and Patil 1984; Myers and Pepin 1990). Further research is needed to assess the robustness of the methods to departures from the delta-lognormal distribution: previous discussion of robustness in this context has tended to focus on point estimation, rather than coverage of confidence intervals (Myers and Pepin 1990; Pennington 1991). If distributional assumptions are a concern, one can use a distribution-free approach, such as non-parametric bootstrapping (Davison and Hinkley 1997) or skewness-correction (Hall 1992).

As stated at the outset, our results will be applicable to the more general situation where we have a delta-lognormal linear model, and wish to calculate a confidence interval for the expected value of the response variable, given the value of one or more explanatory variables.

Acknowledgements I thank an Associate Editor and two anonymous referees for comments that led to improvements in the presentation of the paper. I am grateful to Chris Francis (National Institute of Water & Atmospheric Research, Wellington, New Zealand) for providing the red cod data. Thanks also to Kim Duckworth (New Zealand Ministry of Fisheries) for permission to use the data in this paper.

Appendices

A. Data

The positive observations for the data in the fisheries example are shown below. They are the density (kg/km^2) of red cod (*Pseudophycis bachus*) caught in each of 54 trawls.

10.8	13.2	18.2	19.6	34.2	37.0	41.5	42.3	46.1	46.3	52.7	53.8	55.5	59.2	64.5
66.0	70.2	70.6	74.7	76.8	77.6	78.8	85.0	88.1	89.9	90.8	95.4	100.9	114.1	123.2
131.8	132.7	135.1	141.4	147.4	183.0	223.0	235.3	246.5	253.5	267.1	276.4	293.7	298.6	465.2
584.2	639.2	639.3	663.3	915.7	1004.2	1402.2	1563.2	2948.8						

B. Variance of $\hat{\theta}$

It is convenient in what follows to introduce the random variables $\mathbf{Z} = \{Z_i\}\,(i = 1, \ldots, n)$, defined as

$$Z_i = \begin{cases} 1, & Y_i > 0 \\ 0, & Y_i = 0 \end{cases}$$

Note that $m = \sum_{i=1}^{n} Z_i$ and that we consider repeated sampling under the constraint $m > 1$, i.e. m has a truncated binomial distribution. Then

$$\begin{aligned} \Pr(m = x) &= \frac{1}{1 - (1-\pi)^n - n\pi\,(1-\pi)^{n-1}} \binom{n}{x} \pi^x\,(1-\pi)^{n-x} \\ &= \frac{1}{1-cd} \binom{n}{x} \pi^x\,(1-\pi)^{n-x} \qquad x = 2, 3, \ldots, n \end{aligned}$$

where $c = (1-\pi)^{n-1}$ and $d = 1 + (n-1)\,\pi$ (we suppress the dependence of these two variables on π and n in order to simplify the notation).

The nature of the delta-lognormal model means that the variance of $\hat{\theta}$ (Eq. 6) is given by

$$V\left(\hat{\theta}\right) = V_{\mathbf{Z}}\left\{E\left(\hat{\theta}|\mathbf{Z}\right)\right\} + E_{\mathbf{Z}}\left\{V\left(\hat{\theta}|\mathbf{Z}\right)\right\}.$$

Now

$$E\left(\hat{\theta}\middle|\mathbf{Z}\right) = E\left(\ln p + \bar{x} + \frac{s_X^2}{2}\middle|\mathbf{Z}\right)$$
$$= E\left(\ln p\middle|\mathbf{Z}\right) + E\left(\bar{x}\middle|\mathbf{Z}\right) + E\left(\frac{s_X^2}{2}\middle|\mathbf{Z}\right)$$
$$= \ln p + \mu_X + \frac{\sigma_X^2}{2}$$

In addition, as we are restricting attention to those samples for which $m > 1$, we have conditional independence (given $\mathbf{Z}$) of $\ln p$, $\bar{x}$ and s_X^2, and

$$V\left(\hat{\theta}\middle|\mathbf{Z}\right) = V\left(\ln p + \bar{x} + \frac{s_X^2}{2}\middle|\mathbf{Z}\right)$$
$$= V\left(\ln p\middle|\mathbf{Z}\right) + V\left(\bar{x}\middle|\mathbf{Z}\right) + V\left(\frac{s_X^2}{2}\middle|\mathbf{Z}\right)$$
$$= 0 + \frac{\sigma_X^2}{m} + \frac{\sigma_X^4}{2(m-1)}$$

Thus we have

$$V\left(\hat{\theta}\right) = V_{\mathbf{Z}}\left\{\ln p + \mu_X + \frac{\sigma_X^2}{2}\right\} + E_{\mathbf{Z}}\left\{\frac{\sigma_X^2}{m} + \frac{\sigma_X^4}{2(m-1)}\right\}$$
$$= V_{\mathbf{Z}}(\ln p) + E_{\mathbf{Z}}\left\{\frac{\sigma_X^2}{m} + \frac{\sigma_X^4}{2(m-1)}\right\}$$
$$\approx \frac{V_{\mathbf{Z}}(p)}{E_{\mathbf{Z}}(p)^2} + \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right) + \frac{\sigma_X^4}{2} E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$$
$$= \frac{V_{\mathbf{Z}}(m)}{E_{\mathbf{Z}}(m)^2} + \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right) + \frac{\sigma_X^4}{2} E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$$
$$= \frac{E_{\mathbf{Z}}\left(m^2\right)}{E_{\mathbf{Z}}(m)^2} - 1 + \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right) + \frac{\sigma_X^4}{2} E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$$

where we have used the delta method to obtain an approximation to $V_{\mathbf{Z}}(\ln p)$.

Now

$$E_{\mathbf{Z}}(m) = \frac{1}{1-cd}\sum_{x=2}^{n} x\binom{n}{x}\pi^x(1-\pi)^{n-x}$$

$$= \frac{1}{1-cd}\left\{\sum_{x=0}^{n} x\binom{n}{x}\pi^x(1-\pi)^{n-x} - \sum_{x=0}^{1} x\binom{n}{x}\pi^x(1-\pi)^{n-x}\right\}$$
$$= \frac{n\pi(1-c)}{1-cd}$$

and

$$E_{\mathbf{Z}}\left(m^2\right) = \frac{1}{1-cd}\sum_{x=2}^{n} x^2\binom{n}{x}\pi^x(1-\pi)^{n-x}$$
$$= \frac{1}{1-cd}\left\{\sum_{x=0}^{n} x^2\binom{n}{x}\pi^x(1-\pi)^{n-x} - \sum_{x=0}^{1} x^2\binom{n}{x}\pi^x(1-\pi)^{n-x}\right\}$$
$$= \frac{n\pi(d-c)}{1-cd}$$

Hence

$$V\left(\hat{\theta}\right) \approx \frac{\frac{n\pi(d-c)}{1-cd}}{\left\{\frac{n\pi(1-c)}{1-cd}\right\}^2} - 1 + \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right) + \frac{\sigma_X^4}{2}E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$$
$$= \frac{(d-c)(1-cd)}{n\pi(1-c)^2} - 1 + \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right) + \frac{\sigma_X^4}{2}E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$$

We can obtain unbiased estimates for $\sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right)$ and $\sigma_X^4 E_{\mathbf{Z}}\left(\frac{1}{m-1}\right)$, using the two equations below, which follow directly from normal-theory results for the sample variance:

$$E\left(\frac{s_X^2}{m}\right) = E_{\mathbf{Z}}\left\{E\left(\frac{s_X^2}{m}\middle|\mathbf{Z}\right)\right\} = E_{\mathbf{Z}}\left(\frac{\sigma_X^2}{m}\right) = \sigma_X^2 E_{\mathbf{Z}}\left(\frac{1}{m}\right)$$
$$E\left(\frac{s_X^4}{m+1}\right) = E_{\mathbf{Z}}\left\{E\left(\frac{s_X^4}{m+1}\middle|\mathbf{Z}\right)\right\} = \sigma_X^4 E_{\mathbf{Z}}\left(\frac{1}{m-1}\right).$$

Using these estimates, and estimating π by p (c.f. discussion pp. 6–7), we have an approximately unbiased estimate of $V\left(\hat{\theta}\right)$ given by

$$\hat{V}\left(\hat{\theta}\right) \approx \frac{\left(\hat{d}-\hat{c}\right)\left(1-\hat{c}\hat{d}\right) - m\left(1-\hat{c}\right)^2}{m\left(1-\hat{c}\hat{d}\right)^2} + \frac{s_X^2}{m} + \frac{s_X^4}{2(m+1)}$$

where $\hat{c} = (1-p)^{n-1}$ and $\hat{d} = 1 + (n-1)p$.

C. Profile log-likelihood

We wish to find the values of μ_Y that satisfy

$$w(\mu_Y) = 2\left\{l_p(\hat{\mu}_Y) - l_p(\mu_Y)\right\} = 3.84. \tag{C.1}$$

Now

$$l_p(\hat{\mu}_Y) = l(\hat{\mu}_Y, \hat{\pi}, \hat{\mu}_X; \mathbf{y}) = l(\hat{\pi}, \hat{\mu}_X, \hat{\sigma}_X; \mathbf{y}) \tag{C.2}$$

where $\hat{\pi}$, $\hat{\mu}_X$, $\hat{\sigma}_X$ and $\hat{\mu}_Y$ are maximum likelihood estimates and

$$\begin{aligned} l(\pi, \mu_X, \sigma_X; \mathbf{y}) = k &- \ln\left[1 - (1-\pi)^n - n\pi(1-\pi)^{n-1}\right] \\ &+ m\ln\pi + (n-m)\ln(1-\pi) \\ &- \frac{m}{2}\ln\sigma_X^2 - \frac{(m-1)s_X^2 + m(\bar{x} - \mu_X)^2}{2\sigma_X^2} \end{aligned} \tag{C.3}$$

where k is a constant, and we use the fact that m has the truncated binomial distribution given in Appendix B.

The maximum likelihood estimates of μ_X and σ_X^2 are

$$\hat{\mu}_X = \bar{x} \quad \text{and} \quad \hat{\sigma}_X^2 = \frac{(m-1)}{m}s_X^2.$$

Note that there is no closed-form solution for $\hat{\pi}$ (Finney 1949). Substitution of $\hat{\mu}_X$ and $\hat{\sigma}_X^2$ into Eqs. C.2 and C.3 gives

$$\begin{aligned} l_p(\hat{\mu}_Y) = k' - \frac{m}{2}\ln s_X^2 + \max_{\pi}\Big\{ &- \ln\left[1 - (1-\pi)^n - n\pi(1-\pi)^{n-1}\right] \\ &+ m\ln\pi + (n-m)\ln(1-\pi)\Big\} \end{aligned} \tag{C.4}$$

where k' is a constant, and the maximisation with respect to π is obtained by numerical search.

The second term in $w(\mu_Y)$ is the profile log-likelihood (Eq. 10) and is given by

$$l_p(\mu_Y) = l(\mu_Y, \tilde{\pi}, \tilde{\mu}_X; \mathbf{y}). \tag{C.5}$$

From Eq. 1 we can write

$$\sigma_X^2 = 2(\ln\mu_Y - \ln\pi - \mu_X),$$

and so the likelihood in Eq. C.3 can be rewritten as

$$l(\mu_Y, \pi, \mu_X; \mathbf{y}) = k^* - \ln\left[1 - (1-\pi)^n - n\pi(1-\pi)^{n-1}\right] + m\ln\pi + (n-m)\ln(1-\pi) - \frac{m}{2}\ln(\ln\mu_Y - \ln\pi - \mu_X) - \frac{(m-1)s_X^2 + m(\bar{x} - \mu_X)^2}{4(\ln\mu_Y - \ln\pi - \mu_X)} \quad \text{(C.6)}$$

where k^* is a constant. The profile log-likelihood $l_p(\mu_Y)$ is then found by maximising Eq. C.6 with respect to π and μ_X.

References

Aitchison J (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. J Am Sta Assoc 50:901–908

Berry DA (1987) Logarithmic transformations in ANOVA. Biometrics 43:439–456

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Dennis B, Patil GP (1984) The gamma distribution and weighted multimodal gamma distributions as models of population abundance. Math Biosci 68:187–212

Finney DJ (1941) On the distribution of a variate whose logarithm is normally distributed. J R Stat Soc Ser B 7:155–161

Finney DJ (1949) The truncated binomial distribution. Ann Eugen 14:319–328

Hall P (1992) On removal of skewness by transformation. J Roy Stat Soc Ser B 54:221–228

Land CE (1972) An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics 14:145–158

Lo NCH, Jacobson LD, Squire JL (1992) Indices of relative abundance from fish spotter data based on delta-lognormal models. Can J Fish Aquat Sci 49:2515–2526

Mead R (1988) The design of experiments: statistical principles for practical application. Cambridge University Press

Myers RA, Pepin P (1990) The robustness of lognormal-based estimators of abundance. Biometrics 46:1185–1192

Pennington M (1983) Efficient estimators of abundance, for fish and plankton surveys. Biometrics 39:281–286

Pennington M (1991) On testing the robustness of lognormal-based estimators. Biometrics 47:1623–1624

Smith SJ (1988) Evaluating the efficiency of the Δ-distribution mean estimator. Biometrics 44:485–493

Smith SJ (1990) Use of statistical models for the estimation of abundance from groundfish trawl survey data. Can J Fish Aquat Sci 47:894–903

Stefansson G (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J Mar Sci 53:577–588

Venzon DJ, Moolgavkar SH (1988) A method for computing profile-likelihood-based confidence intervals. Appl Stat 37:87–94

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol Modell 88:297–308

Author Biography

David Fletcher is an Associate Professor in the Department of Mathematics and Statistics at the University of Otago, Dunedin, New Zealand. He is also Co-Director of Proteus Wildlife Research Consultants, which specialises in statistical ecology. He has collaborated extensively with zoologists, both at Otago University and in the New Zealand Department of Conservation.