Estimation for Delta-Lognormal Distribution Review of Estimation Methods for Parameters of the Delta-Lognormal Distribution 1 Estimation for Delta-Lognormal Distribution

Preprint · June 2019

DOI: 10.13140/RG.2.2.13504.38408

Mary C. Christman
MCC Statistical Consulting LLC

June 9, 2019

Review of Estimation Methods for Parameters of the Delta-Lognormal Distribution[1]

Mary C. Christman
MCC Statistical Consulting LLC,
Gainesville, FL 32605,
marycchristman@gmail.com

[1] Copyright © 2019 Mary C Christman, MCC Statistical Consulting, All rights reserved.

Table of Contents

Table of Figures

Table of Tables

Introduction

In this report we review the current approaches for estimating the parameters of the mixture distribution known as the delta-lognormal distribution. In cases where we discuss the more general form of the distribution, we consider only those distributions where the non-zero data are strictly positive, and the degenerate distribution is at $x = 0$.

The review was undertaken for several reasons not the least of which was the need to summarize in one location, the current methods used to estimate the parameters of the component distributions as well as the combined estimator of the population mean or total. A related estimation requirement is the need for a statement of the variance of the population mean (total) and estimators for this variance.

There has been quite a large volume of literature in which the delta-lognormal distribution is used to estimate means and totals in the fisheries literature (*cf.* Lo et al., 1992; Stefansson, 1996; Rodrigues – Martin et al., 2003; IATTC, 2007; Ingram et al., 2008; Cass-Calay 2010; Ingram, 2011). Unfortunately in some instances either the approach was ad hoc estimation of means, variances and confidence intervals or the methods were applied without regard to the appropriateness of the model for the population of interest.

Throughout, we assume that data have been collected either by simple random sampling or at least that the data are independent, i.e. the data are not collected such that observations are correlated, and the data are representative of the population, i.e. every unit in the population has a positive probability of being selected to be in the sample.

We start by describing the distribution and its moments. Then we cover point and confidence interval estimation under simple random sampling and under estimation using a zero-altered model with covariates. Finally we review two possible failures of the assumptions of the model, namely whether the estimators of the components of the model are independent and failure of the assumption of lognormality of the non-zero data. Simulations are done to compare different estimators of the means and variances, coverage of different confidence interval estimators, and failures of the assumptions.

Future work includes estimation under complex survey design and hypothesis testing.

Delta – Distribution & Moments

The delta distribution is a mixture of a degenerate distribution at $X = 0$[2] and a conditional probability distribution for X given $X > 0$:

$$\begin{aligned} &Prob(X = 0) &&= \theta, && x = 0 \\ &Prob(X \subset \{x, x + dx\}; X > 0) &&= h(x)dx, && x > 0 \end{aligned} \qquad (1)$$

where $h(x)$ is the probability density function of the non-zero data. Let the mean and variance of the non-zero distribution be

$$E[X; X > 0] = \alpha$$

and

$$Var[X; X > 0] = \beta.$$

Then, the mean and variance of the delta–distribution are

$$E[X] = \gamma = (1 - \theta)\alpha$$

and

$$Var[X] = \delta = (1 - \theta)\beta + \theta(1 - \theta)\alpha^2$$

(Aitchison, 1955) under the assumption that the expectation $E[X; X > 0, \theta] = E[X; X > 0]$ (and $Var[X; X > 0, \theta] = Var[X; X > 0]$) , i.e. the mean (and variance) of the non-zero distribution does not depend on the proportion of non-zeroes, (equivalently, it does not depend on the Bernoulli distribution of absence ($X = 0$) or presence ($X > 0$)). Note that as θ approaches 0, $E[X] \to \alpha$ and $Var[X] \to \beta$, whereas as $\theta \to 1$, both $E[X]$ and $Var[X] \to 0$.

Delta – Lognormal Distribution

The lognormal distribution is defined as

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\left(\frac{1}{x}\right)\exp\left(\frac{-(\log(x) - \mu)^2}{2\sigma^2}\right), \quad x > 0, \; \mu \in (-\infty, +\infty), \; \sigma^2 > 0$$

such that $Y = \log(X) \sim N(\alpha, \beta)$. The mean[3] and variance of X then is

[2] In fact, it could be for a constant ≠ 0. If that is the case and the alternative distribution includes the constant in the sample space, then the definition and moments need to be appropriately modified. In this review, we discuss only the case given in the text.

$$E_{LN}[X; X > 0] = \alpha = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

and

$$Var_{LN}[X; X > 0] = \beta = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

(Casella & Berger, 2002). Finney (1941) derived unbiased minimum variance estimators of the two parameters α and β.

If the lognormal distribution is used for the non-zero component in the delta–distribution, the mean and variance of the mixture distribution then is:

$$E[X] = \gamma = (1 - \theta) \exp\left(\mu + \frac{\sigma^2}{2}\right) \qquad (2)$$

and

$$Var[X] = \delta = (1 - \theta) \exp(2\mu + \sigma^2) [\exp(\sigma^2) - (1 - \theta)] \qquad (3)$$

(Aitchison, 1955).

Estimators Based on Simple Random Sampling

Estimation of the First Two Moments of the Delta-Lognormal Distribution

There are several approaches to estimating the mean and variance of a random variable from a mixture of distributions. To show these, let us suppose we took a simple random sample of size n from this mixture and observe r zeroes and $m = (n - r)$ non-zeroes.

Naïve Approach

First is an approach in which the parameters $\boldsymbol{\varphi} = (\theta, \mu, \sigma^2)$ in the mean and variance of X are replaced with sample estimators $\widehat{\boldsymbol{\varphi}} = \left(\frac{r}{n}, \bar{y}, s_y^2\right)$ so that

[3] The median of the lognormal distribution is $Median = \exp(\mu) < Mean = \exp\left(\mu + \frac{\sigma^2}{2}\right)$. We mention this since a common approach is to simply back-transform the logged data by exponentiation but the resulting estimate is biased low for the mean of the distribution.

$$\widehat{E[X]} = \hat{\gamma}_{Nai} = (1-\hat{\theta})\exp\left(\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right) = \left(1-\frac{r}{n}\right)\exp\left(\bar{y}+\frac{s_y^2}{2}\right)$$

and

$$\widehat{Var[X]} = \hat{\delta}_{Nai} = \left(1-\frac{r}{n}\right)\exp\left(2\bar{y}+s_y^2\right)\left[\exp\left(s_y^2\right)-\left(1-\frac{r}{n}\right)\right]$$

where

$$\bar{y} = \frac{1}{(n-r)}\sum\nolimits_{i=1}^{n-r}\log(x_i)$$

is the sample mean of the natural log-transformed non-zero observations, and

$$s_y^2 = \frac{1}{(n-r-1)}\sum\nolimits_{i=1}^{n-r}(\log(x_i)-\bar{y})^2$$

is the sample variance of the natural log-transformed non-zero observations. The estimators, $\hat{\gamma}_{Nai}$ and $\hat{\delta}_{Nai}$, tend to be biased high and the size of the bias increases as the variance of X (i.e. β) increases. The bias decreases slowly as the sample size increases. Hence the naïve estimators are not recommended for small sample sizes. Interestingly Zou *et al*. (2009) use this estimator of the mean to develop a confidence interval estimator that has excellent coverage relative to many other interval estimators (see Section on confidence interval estimation).

General Unbiased Estimators (UE)

In order to obtain unbiased estimators of the parameters γ and δ, we require unbiased estimators of $(\theta, \alpha, \alpha^2, \beta)$. The sample estimate $\hat{\theta} = \frac{r}{n}$ is sufficient and unbiased for θ (Casella & Berger, 2002). Let $a_{(n-r)}$, $b_{(n-r)}$, and $c_{(n-r)}$ be sufficient unbiased estimators of α, α^2, and β, respectively. Then, minimum variance, unbiased estimators of $E[X]$ and $Var[X]$ are given by

$$\widehat{E[X]} = \hat{\gamma} = (1-\hat{\theta})a_{(n-r)} \tag{1}$$

and

$$\widehat{Var[X]} = \hat{\delta} = \left(1-\hat{\theta}\right)c_{(n-r)} + \hat{\theta}\left(1-\hat{\theta}\right)b_{(n-r)} \tag{2}$$

(Aitchison, 1955). These estimators are said to be UMVUE: uniformly minimum variance unbiased estimators.

When There Are No Assumptions on the Distribution of the Non-Zero Data

If no assumptions are placed on the distribution of the non-zero data, an unbiased estimator of the mean (α) of the non-zero data on the original data scale is the sample mean of the non-zero values:

$$\hat{\alpha} = a_{(n-r)} = \frac{1}{(n-r)}\sum_{i=1}^{n-r} x_i.$$

Similarly, an unbiased estimator of the variance of the non-zero data (β) is

$$\hat{\beta} = c_{(n-r)} = \frac{1}{(n-r-1)}\sum_{i=1}^{n-r} (x_i - a_{(n-r)})^2.$$

Finally, an unbiased estimator of α^2 is

$$\widehat{\alpha^2} = b_{(n-r)} = \hat{E}[X^2; X > 0] - \widehat{Var}[X; X > 0] = a_{(n-r)}^2 - \frac{c_{(n-r)}}{(n-r)} \tag{3}$$

(Glasser, 1960). Then, the delta-estimators reduce to the more common sample mean and variance:

$$\widehat{E[X]} = \hat{\gamma} = \left(1 - \frac{r}{n}\right)\frac{1}{(n-r)}\sum_{i=1}^{n-r} x_i = \bar{x}$$

and

$$\widehat{Var[X]} = \hat{\delta} = \left(1 - \frac{r}{n}\right)\left\{c_{(n-r)} + \frac{r}{n}\left(a_{(n-r)}^2 - \frac{c_{(n-r)}}{(n-r)}\right)\right\} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = s_x^2.$$

When the Log-normal Distribution is Assumed for the Non-Zero Data

To unbiasedly estimate $E[X]$ and $Var[X]$ under the assumption that the non-zero data are lognormally distributed requires unbiased estimators of the lognormal mean (α), the mean squared (α^2) and the variance (β). Finney (1941) derived these and proposed the following estimators of the mean and variance of the non-zero data:

$$\widehat{E_{LN}}[X; X > 0] = \hat{\alpha} = \exp(\bar{y})\, g_{(n-r)}\left(\frac{1}{2}s_y^2\right) \tag{4}$$

and

$$\widehat{Var_{LN}}[X; X > 0] = \hat{\beta} = \exp(2\bar{y})\left\{g_{(n-r)}(2s_y^2) - g_{(n-r)}\left(\frac{n-r-2}{n-r-1}s_y^2\right)\right\} \tag{5}$$

where $\bar{y}$ and s_y^2 are as described above for the naïve estimator and

$$g_m(t) = 1 + \frac{m-1}{m}t + \sum_{j=2}^{\infty} \frac{(m-1)^{2j-1}}{j!\prod_{i=2}^{j}(m+2i-3)}\left(\frac{t}{m}\right)^j$$

is a generalized hypergeometric series such that

$$E\left[g_m\left(as_y^2\right)\right] = \exp\left(a\left(\frac{m-1}{m}\right)\sigma^2\right)$$

(see also Aitchison, 1955; Smith, 1988; Bradu & Mundlak, 1970). Aitchison (1955) used Finney's results to obtain unbiased estimators of $E[X]$ and $Var[X]$ of the mixture distribution:

$$\widehat{E[X]} = \hat{\gamma}_{Ait} = \begin{cases} \left(1-\hat{\theta}\right)\exp(\bar{y})\, g_{(n-r)}\left(\frac{1}{2}s_y^2\right), & r < n \\ \frac{x}{n}, & r = n-1 \\ 0, & r = n \end{cases} \qquad (6)$$

and

$$\widehat{Var}[X] = \hat{\delta}_{Ait}$$
$$= \begin{cases} \left(1-\hat{\theta}\right)\exp(2\bar{y})\left\{g_{(n-r)}\left(2s_y^2\right) - \left(1-\frac{r}{n-1}\right)g_{(n-r)}\left(\frac{n-r-2}{n-r-1}s_y^2\right)\right\}, & r < n \\ \left(\frac{x}{n}\right)^2, & r = n-1 \\ 0, & r = n \end{cases} \qquad (7)$$

These estimators are minimum variance estimators as well as unbiased and hence have the same or lower variance than any other linear unbiased estimator of the respective parameter. Owen and DeRouen (1980) showed that the minimum variance unbiased estimator of γ given by Aitchison (1955) was slightly better than the estimators they obtained, when compared on the basis of mean squared error.

Variance of the Estimated Mean $\widehat{E[X]} = \hat{\gamma}_{Ait}$

Of interest in most studies is not only estimation of the mean and variance of the distribution but more importantly, the variance of the estimated mean. We first describe the variance of $\widehat{E[X]} = \hat{\gamma}_{Ait}$ and then its estimation.

Aitchison and Brown (1957) obtained an approximate variance for the estimated mean of the mixture distribution assuming large sample size, n, and the proportion of zeroes in the sample ($\hat{\theta}$) not close to 1:

$$Var_{asy}[\hat{\gamma}_{Ait}] = \frac{\exp(2\mu + \sigma^2)}{n}\left[\theta(1-\theta) + \frac{1}{2}(1-\theta)(2\sigma^2 + \sigma^4)\right].$$

Owen and DeRouen (1980) used a naïve approximation of this variance by replacing the parameters with sample estimates: $\hat{\mu} = \bar{y}$, $\hat{\sigma}^2 = s_y^2$, and $\hat{\theta} = \frac{r}{n}$:

$$\widehat{Var}_{asy}[\hat{\gamma}_{Ait}] = \frac{\exp\left(2\bar{y} + s_y^2\right)}{n}\left[\frac{r}{n}\left(1-\frac{r}{n}\right) + \frac{1}{2}\left(1-\frac{r}{n}\right)\left(2s_y^2 + (s_y^2)^2\right)\right].$$

They showed that the estimator $\widehat{Var}_{asy}[\hat{\gamma}_{Ait}]$ is virtually identical to $Var_{asy}[\hat{\gamma}_{Ait}]$ for large n, i.e. $n > 15$. Unfortunately, they considered only whether their estimator is a reasonable estimator of the quantity being estimated; they did not address whether $Var_{asy}[\hat{\gamma}_{Ait}]$ is a good estimator for $Var[\hat{\gamma}_{Ait}]$. In fact, it is not a good estimator based on a simulation study we describe later.

Smith (1988) on the other hand, derived an exact form for the variance $Var[\hat{\gamma}_{Ait}]$ of the estimated mixture mean. First, Smith obtained an alternative form for the generalized hypergeometric series, $g_m(t)$, used in unbiased estimators of $\exp(k\sigma^2)$:

$$\Phi_m(t) = \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m}{2}+j\right)} \frac{1}{j!}\left(\frac{t}{4}\right)^j$$

where $\Gamma(q)$ is the gamma function $\Gamma(q) = \int_0^\infty x^{q-1}e^{-x}dx$. The mean of $\Phi_m(t)$ for $t = \frac{(m-1)^2}{m} s_y^2$ is

$$E\left[\Phi_m\left(\frac{(m-1)^2}{m}s_y^2\right)\right] = \exp\left(\left(\frac{m-1}{m}\right)\frac{\sigma^2}{2}\right).$$

Smith uses this function to obtain an exact variance for the estimator $\hat{\gamma}$:

$$Var_{ex}[\hat{\gamma}_{Ait}] = \exp(2\mu + \sigma^2)\left[\sum_{i=1}^{n}\left\{\Pr(n-r=i)\left(\frac{i}{n}\right)^2 \exp\left(\frac{\sigma^2}{i}\right)\Phi_m\left(\frac{i-1}{i}\sigma^4\right)\right\} - (1-\theta)^2\right].$$

Using this form of the variance of the estimator, Smith showed that the efficiency of the variance of the estimator of the mean depends on the sample size, n, the proportion of zeros in the sample, θ, and the variance of Y (σ^2). Note that it is also a function of the mean and so as the mean increases, so does the variance. Increases in n increase efficiency while increases in either θ or σ^2 decrease efficiency. Unfortunately, Smith (1988) did not provide an estimator for $Var_{ex}[\hat{\gamma}_{Ait}]$.

Hoyle (1968) derived an unbiased estimator of the variance of the estimator of $E_{LN}[X; X > 0]$, $\hat{\alpha} = \exp(\bar{y})\, g_{(n-r)}\left(\frac{1}{2} s_y^2\right)$, using the $g_m(t)$ function described by Finney (1941):

$$\widehat{Var}[\hat{\alpha}] = \exp(2\bar{y}) \left\{ g_{(n-r)}^2 \left(\frac{1}{2} s_y^2\right) - g_{(n-r)} \left(\frac{n-r-2}{n-r-1} s_y^2\right) \right\}.$$

Pennington (1983) then used Hoyle's results to derive an estimator for $Var_{ex}[\hat{\gamma}]$, namely

$$\widehat{Var_{Pen}}[\hat{\gamma}_{Ait}] = \begin{cases} \left(1 - \frac{r}{n}\right) \exp(2\bar{y}) \left\{ \left(1 - \frac{r}{n}\right) g_{(n-r)}^2 \left(\frac{1}{2} s_y^2\right) - \left(1 - \frac{r}{n-1}\right) g_{(n-r)} \left(\frac{n-r-2}{n-r-1} s_y^2\right) \right\}, & r < n \\ \left(\frac{x}{n}\right)^2, & r = n-1 \\ 0, & r = n \end{cases}$$

which is now widely used for estimating abundance indices in fisheries research.

Simulation Study

In order to determine the behavior of the different estimators of the population means, population variances and variances of the means, we ran a simulation study. The simulations were performed for two sample sizes, $n = 50$ and 100, and four combinations of means and variances for the lognormal distribution (Table 1). The expected proportion of non-zeroes was fixed at 0.4 for these simulations.

Table 1. Parameters used in simulations. The expected proportion of non-zeroes was set to $(1 - \theta) = 0.4$.

$E[\log(X)] = \mu$	$Var[\log(X)] = \sigma^2$	$E[X; X > 0] = \alpha$	$Var[X; X > 0] = \beta$	$E[X] = \gamma$	$Var[X] = \delta$
1	1	4.482	34.513	1.793	18.63
1	4	20.086	21623.040	8.034	8746.04
2	1	12.182	255.016	4.873	137.63
2	4	54.598	159773.800	21.839	64624.96

For each combination of sample size and parameter values, 100,000 simulations were drawn from a delta-lognormal distribution as follows:

1) A random number of non-zero observations (m) in a sample of size n was selected from a binomial distribution with $n = 50$ or 100 and $(1 - \theta) = 0.4$.
2) Given m, m non-zero observations were randomly drawn from a lognormal distribution where $\mathrm{Y} = \log(X) \sim N(\mu, \sigma^2)$ and $\{\mu, \sigma^2\}$ were as given in Table 1.

The sample then is composed of the m non-zero values and $r = n - m$ zeroes where m is random. For each simulated sample, we calculated the following statistics:

1) $\bar{x}$, the sample mean on the original scale,
2) s_x^2, the sample variance on the original scale,
3) $\hat{\gamma}_{Ait}$, Aitchison's (1955) estimator of the mean,
4) $\hat{\delta}_{Ait}$, Aitchison's (1955) estimator of the variance of X,
5) $\hat{\gamma}_{Nai}$, the naïve estimator of the mean of X
6) $\hat{\delta}_{Nai}$, the naïve estimator of the variance of X,
7) $\widehat{Var_{Pen}}[\hat{\gamma}_{Ait}]$, Pennington's (1983) estimator of the variance of Aitchison's estimator $\hat{\gamma}$, and,
8) $\widehat{Var}_{asy}[\hat{\gamma}_{Ait}]$, Owen and DeRouen's (1980) estimator of the asymptotic approximate variance described by Aitchison (1955).

Simulations were summarized by averaging the 100,000 simulated values of each statistic (Table 2).

As expected, the sample mean ($\bar{x}$) and Aitchison's estimator ($\hat{\gamma}_{Ait}$) are unbiased for the mean of X. Conversely, the naïve estimator ($\hat{\gamma}_{Nai}$) of the mean of the distribution is biased. The next comparison is among the three estimators for the variance of X. Both s_x^2 and $\hat{\delta}_{Ait}$ have very similar values for all combinations of parameters and sample sizes, implying that either is acceptable as an estimator of $V[X]$. Not surprisingly, the naïve estimator of $Var[X]$, i.e. $\hat{\delta}_{Nai}$, behaves very poorly; it is highly biased being over twice as large as the true variance.

Two estimators of the variance of the sample mean, $Var(\bar{x})$, were calculated: $\frac{s_x^2}{n}$ and $\frac{\hat{\delta}_{Ait}}{n}$, the second recommended by Aitchison (1955). Both are unbiased (Table 2), as expected, but s_x^2 is slightly more variable from sample to sample (not shown) than $\hat{\delta}_{Ait}$ for smaller sample sizes. So, for lognormal distributions with large variances, $\hat{\delta}_{Ait}/n$ might be preferred as an estimator of the variance of $\bar{x}$ when the distribution of the non-zero data is lognormal (see section on failure of the assumptions).

In order to determine the behavior of the estimators of the variance of Aitchison's estimator $\hat{\gamma}_{Ait}$(eq. 6), namely $\widehat{Var}_{Pen}[\hat{\gamma}_{Ait}]$ and $\widehat{Var}_{asy}[\hat{\gamma}_{Ait}]$, we estimated the true variance of $\hat{\gamma}_{Ait}$ by calculating the variance of the 100,000 simulated values of $\hat{\gamma}_{Ait}$. This is a reasonably accurate approximation[4] of the true variance given by Smith (1988). The Pennington estimator of the variance of Aitchison's estimator $\hat{\gamma}_{Ait}$, $\widehat{Var}_{Pen}[\hat{\gamma}_{Ait}]$, is, as expected, unbiased for the true variance (Table 2). Conversely, the asymptotic estimator proposed by Aitchison and Brown (1957) and promoted by several authors in applications is a poor estimator in all cases except when the lognormal variance is small. Even then, it tends to be biased high but the bias is small. It is a very poor estimator when the variance of the lognormal distribution is large.

A comparison of $V(\bar{x})$ and $V(\hat{\gamma}_{Ait})$ supports the conclusions from Aitchison (1955) and Smith (1988) indicating that for most cases, $\hat{\gamma}$ is a more efficient estimator of the mean of the mixture than $\bar{x}$. Note though that the variance of the sample mean is very similar to that of Aitchison's estimator when the variance of the lognormal distribution is small, which implies that there is little gain in efficiency if the non-zero data are moderately skewed.

[4] We assume that 100,000 simulations of samples from the delta-lognormal distribution is sufficient to remove or greatly reduce the Monte Carlo error from simulating rather than calculating the variance directly using Smith's equation.

Table 2. Estimators of the mean of the mixture and their variances for various combinations of means and variances of the mixture distribution and sample sizes. Also shown are alternative estimators of the variance of the estimators of the mean. In all cases, the mixture distributions were composed of an expected 60% zeroes and non-zeroes distributed according to a lognormal distribution with varying means and variances. E[X] and V[X] are is the mean and variance, respectively, of the mixture distribution.

Sample Size	Parameter Values		Estimates of $E[X]$			Estimates of $V[X]$		
	$E[X]$	$V[X]$	$\bar{x}$	$\hat{\gamma}_{Ait}$	$\hat{\gamma}_{Nai}$	$\hat{\delta}_{Nai}$	s_x^2	$\hat{\delta}_{Ait}$
50	1.79	18.63	1.79	1.79	1.87	34.51	18.65	18.65
50	8.03	8746.00	8.06	8.03	11.41	21623.04	7605.13	8628.27
50	4.87	137.63	4.87	4.87	5.07	255.02	139.85	136.95
50	21.83	64625.00	21.76	21.86	31.06	159773.80	55660.0	70395.0
100	1.79	18.63	1.79	1.79	1.83	22.12	18.70	18.60
100	8.03	8746.00	8.08	8.04	9.45	112151.6	8470.0	8720.0
100	4.87	137.63	4.88	4.88	4.97	163.84	137.90	137.90
100	21.83	64625.00	21.83	21.81	25.67	1155249.0	54093.7	65846.8

Sample Size	Parameter Values		Parameter Value	Estimates of $V[\bar{x}]$		Parameter Value	Estimates of $V[\hat{\gamma}_{Ait}]$	
	$E[X]$	$V[X]$	$V[\bar{x}]$	$\frac{s_x^2}{n}$	$\frac{\hat{\delta}_{Ait}}{n}$	$V[\hat{\gamma}_{Ait}]$	$\hat{V}_{Pen}[\hat{\gamma}_{Ait}]$	$\hat{V}_{asy}[\hat{\gamma}_{Ait}]$
50	1.79	18.63	0.373	0.373	0.373	0.343	0.343	0.452
50	8.03	8746.00	151.9	152.1	171.8	48.84	49.52	682.6
50	4.87	137.63	2.7	2.8	2.7	2.531	2.517	3.323
50	21.83	64625.00	1292.5[*]	1217.1[*]	1298.2[*]	396.28	394.62	6107.60
100	1.79	18.63	0.187	0.187	0.186	0.17	0.17	0.19
100	8.03	8746.00	87.5	84.7	87.2	23.2	23.10	60.96
100	4.87	137.63	1.376	1.379	1.379	1.245	1.26	1.43
100	21.83	64625.00	646.2	643.9	645.3	171.97	171.57	455.14

[*] Based on 500,000 simulations

Confidence Interval Estimation of the Mean of the Mixture Distribution

Aitchison's estimator $\hat{\gamma}_{Ait} = \left(1-\hat{\theta}\right)\exp(\bar{y})\, g_{(n-r)}\left(\frac{1}{2}s_y^2\right)$ is the product of functions of three different statistics, each of which has an unknown distribution. Hence, the construction of a confidence interval for the population mean using $\hat{\gamma}_{Ait}$ must rely either on asymptotic arguments, knowledge of the true distribution of the estimator $\hat{\gamma}_{Ait}$ or on alternative methods for obtaining an approximate distribution for the estimator, e.g. bootstrapping (Efron & Tibshirani, 1986). An additional problem that complicates obtaining good coverage of confidence intervals arises because the estimator of the mean is positively correlated with the estimator of the variance and hence with the estimated standard error of the mean. As a result, estimated means that are lower than the true mean will also have confidence intervals that are narrower than those associated with estimated means larger than the true mean. As a result, tail coverages for a symmetric confidence interval (e.g. $\hat{\tau} \pm 1.96\hat{\sigma}_{\hat{\tau}}$) are unequal. The correlation between estimated means and variances is not uncommon in skewed distributions (*cf*. Gregoire & Schabenberger, 1999).

Asymptotic arguments rely on the central limit theorem but the speed of convergence to a normal distribution is dependent here on not only the sample size but also the probability of observing a zero. To determine the shape of the distribution of Aitchison's estimator of the mean for small to intermediate sized samples and the effect of the parameter values of the lognormal distribution on that shape, we ran a second simulation study.

We simulated data from a mixture of a Binomial and lognormal distributions using

a) 3 sample sizes (25, 50 and 100),
b) 3 probabilities of observing a non-zero observation (0.1, 0.3, and 0.6), and
c) 2 different set of parameters for the lognormal distribution ($\{\mu, \sigma^2\} = \{1,1\}$ and $\{1,4\}$).

For each simulated dataset, we calculated Aitchison's estimator of the mean ($\hat{\gamma}_{Ait}$) and Pennington's estimator of the variance of the estimated mean ($\hat{V}_{Pen}\left[\hat{\gamma}_{Ait}\right]$). The correlation between the estimated mean and the standard error of the mean was approximately 0.88 – 0.90 for all samples sizes and parameter values (not shown). Figures 1 and 2 show the histograms for 100 simulated means based on the Aitchison estimator for the various samples sizes and parameter values. It has been suggested that the estimator $\hat{\gamma}_{Ait}$ itself might be log-normally distributed (Shono 2008) so we tested each simulated distribution for lognormality. As expected every test rejected the null hypothesis of normality of the $\log(\hat{\gamma}_{Ait})$ ($p > 0.15$, results not shown).

Approaches to Confidence Interval Estimation

Asymptotic Theory

Under the assumption that the sample size is sufficiently large to invoke the Central Limit Theorem (Casella & Berger, 2002), one possible $(1-\alpha)100\%$ confidence interval estimator for $E[X] = \gamma$ is

$$\widehat{E[X]} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{E[X]})}$$

where $\widehat{E[X]}$ is an estimator of the mean γ of the mixture distribution, $z_{\alpha/2}$ is the critical value for area $\alpha/2$ in the upper tail of a standard normal distribution, and $\widehat{Var}(\widehat{E[X]})$ is the estimator of the variance of the estimated mean.

Transformation + Asymptotic Theory

Shono (2008) suggested an alternative confidence interval estimator based on a natural log-transformation of the estimated mean, calculation of a confidence interval assuming asymptotic normality of the transformed estimator, and a back-transformation of the endpoints of confidence interval. His exact equations are based on generalized linear models for both the proportion of zeroes and the mean of the non-zero observations. We can borrow the idea though for the simpler case that uses Aitchison's estimator, $\hat{\gamma}_{Ait}$, and Pennington's estimator of the variance of Aitchison's estimator, $\widehat{Var}_{Pen}[\hat{\gamma}_{Ait}]$.

One first constructs the symmetric interval:

$$\log(\hat{\gamma}_{Ait}) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\gamma}_{Ait}))}$$

where $Var(\log(\hat{\gamma}_{Ait}))$ is derived using the Delta Method (Casella & Berger, 2002, pg. 240) and then estimated by replacing the unknown parameters with their estimates:

$$\widehat{Var}(\log(\hat{\gamma}_{Ait})) \approx \frac{\widehat{Var}_{Pen}(\hat{\gamma}_{Ait})}{\hat{\gamma}_{Ait}{}^{2}}.$$

A $(1-\alpha)100\%$ confidence interval for γ on the original data scale uses the back-transformed endpoints of the symmetric interval:

$$\left\{\hat{\gamma}_{Ait}\exp^{-1}\left(z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\gamma}_{Ait}))}\right),\ \hat{\gamma}_{Ait}\exp\left(z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\gamma}_{Ait}))}\right)\right\}.$$

This interval will be asymmetric but presumably should have better coverage than the one based on asymptotic theory alone if the distribution of $\hat{\gamma}_{Ait}$ is approximately lognormal.

An alternative to the natural-logarithm transformation would be to use the square root transformation. In that case, the approximate variance estimator, based on the Delta Method, would be $\widehat{Var}\left(\sqrt{\hat{\gamma}_{Ait}}\right) \approx \widehat{Var}_{Pen}(\hat{\gamma}_{Ait})/4\,\hat{\gamma}_{Ait}$ and the back-transformation would be to square the two endpoints of the interval:

$$\sqrt{\hat{\gamma}_{Ait}} \pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\sqrt{\hat{\gamma}_{Ait}}\right)}.$$

Both transformations suffer when a sample contains all zeroes as could occur for small sample sizes and large probabilities of observing a 0. In those cases, the variance of the transformed estimated mean cannot be calculated.

Bootstrapping

*S*ince Aitchison's estimator is unbiased and the distribution of the estimator is unknown, one might consider using a form of the percentile bootstrap confidence interval method (Efron & Tibshirani, 1986) to obtain a $(1-\alpha)100\%$ confidence interval for γ. There are several bootstrapping alternatives that could be taken. The most simplistic is to use the percentiles of the distribution of the bootstrapped estimates $\{\hat{\gamma}_b; b = 1, \dots, B\}$ from the B bootstrap samples. As will be seen, this method is not recommended.

A second method is the bootstrap-t method in which the bootstrap estimates are t-statistics, i.e. of the form $t_b = \frac{\hat{\gamma}_b - \hat{E}[\hat{\gamma}]}{\hat{V}_b[\hat{\gamma}]}$. This approach normalizes the bootstrap $\hat{\gamma}_b$ estimates to the same mean and variance unlike the simpler percentile method approach of using the $\hat{\gamma}_b$directly. One uses the sample estimate as the estimate of the mean, that is, $\hat{\gamma}_{Ait} = \hat{E}[\hat{\gamma}_{Ait}]$ and uses the Pennington estimator of the variance of $\hat{\gamma}_{Ait}$, $\hat{V}_{Pen}[\hat{\gamma}_{Ait}]$, for the denominator of the t-statistic.

An alternative, which has been proposed (Efron & Tibshirani, 1993; Gregoire & Shabenberger, 1999) as a method for cases where the variance is a function of the mean, is nested bootstrapping. In this case, the variance in the denominator of t_b is not estimated by a known function such as Pennington's estimator applied to each bootstrap sample but is instead calculated by a nested resampling of the b^{th} bootstrap sample to obtain a measure of variability of $\hat{\gamma}_b$ given the sample. One such approach is to use a leave-one-out jackknife method at the second level of resampling (Efron & Tibshirani, 1993).

Comment 1: When using bootstrapping it is important that the bootstrap samples be selected according to the sampling design used to obtain the original sample. For example, if the population is finite and a sample was taken randomly without replacement, then the bootstrap sample should be selected accordingly (McCarthy & Snowden, 1985; Rao & Wu, 1988; Booth et al. 1994; Sitter 1992a, 1992b; Mohammadi et al., 2014).

Comment 2: Bootstrapping should only be performed if the researcher believes that the sample is a reasonable representation of the variability within the population from which it was taken. This usually requires fairly large sample sizes for highly skewed data.

MOVER technique of Zou et al. 2009

Zou et al. (2009) use an approach based on the naïve estimator $\hat{\gamma}_{Nai} = \left(1-\frac{r}{n}\right)\exp\left(\bar{y}+\frac{s_y^2}{2}\right)$ that they refer to as the Method Of Variance Estimates Recovery (MOVER). The method involves several steps as outlined below.

Of interest is estimating a $(1-\alpha)100\%$ confidence interval for $\gamma = (1-\theta)\exp\left(\mu+\frac{\sigma^2}{2}\right)$. The estimation method starts by considering instead interval estimation for $\log(\gamma) = \log(1-\theta) + \left(\mu+\frac{\sigma^2}{2}\right) = \beta_1 + \beta_2$. The algorithm is as follows:

1) Calculate an approximate $(1-\alpha)100\%$ confidence interval for θ (Agresti & Coull, 1998):

$$\frac{\left[\hat{\theta} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\left\{\hat{\theta}\left(1-\hat{\theta}\right) + \frac{z_{\alpha/2}^2}{4n}\right\}}\right]}{1+\frac{z_{\alpha/2}^2}{n}}$$

   where n is the sample size, r is the number of zeroes in the sample, and $\hat{\theta} = \frac{r}{n}$;

2) Calculate the approximate $(1-\alpha)100\%$ confidence interval for $(1-\theta)$ as simply $\{1-LB_\theta, 1-UB_\theta\}$ where LB_θ (UB_θ) is the lower (upper) bound of the interval in step 1;

3) Calculate the approximate endpoints for a confidence interval for $\log(1-\theta) = \beta_1$ as

$$l_{\beta_1} = \log(1-UB_\theta)$$
$$u_{\beta_1} = \log(1-LB_\theta);$$

4) Calculate an approximate $(1-\alpha)100\%$ confidence interval for $\beta_2 = \left(\mu+\frac{\sigma^2}{2}\right)$:

$$l_{\beta_2} = \bar{y} + \frac{s_y^2}{2} - \sqrt{z_{\alpha/2}^2\frac{s_y^2}{(n-r)} + \left\{\frac{s_y^2}{2}\left(1-\frac{(n-r-1)}{\chi^2_{1-\frac{\alpha}{2},(n-r-1)}}\right)\right\}^2}$$

$$u_{\beta_2} = \bar{y} + \frac{s_y^2}{2} + \sqrt{z_{\alpha/2}^2 \frac{s_y^2}{(n-r)} + \left\{\frac{s_y^2}{2}\left(\frac{(n-r-1)}{\chi^2_{\frac{\alpha}{2},(n-r-1)}} - 1\right)\right\}^2}$$

where $\chi^2_{a,b}$ is the critical value for right tail probability a on b degrees of freedom.

5) The estimators of the $(1-\alpha)100\%$ confidence interval endpoints for $\log(\gamma) = \beta_1 + \beta_2$ are given by

$$L = \log\left(1-\hat{\theta}\right) + \bar{y} + \frac{s_y^2}{2} - \sqrt{\left(\log\left(1-\hat{\theta}\right) - l_{\beta_1}\right)^2 + \left(\bar{y} + \frac{s_y^2}{2} - l_{\beta_2}\right)^2}$$

$$U = \log\left(1-\hat{\theta}\right) + \bar{y} + \frac{s_y^2}{2} + \sqrt{\left(u_{\beta_1} - \log\left(1-\hat{\theta}\right)\right)^2 + \left(u_{\beta_2} - \bar{y} - \frac{s_y^2}{2}\right)^2}.$$

6) The approximate $(1-\alpha)100\%$ confidence interval for γ is $\{\exp(L), \exp(U)\}$.

Comparison of Methods

We ran a simulation study to compare the coverage rates for the interval estimation procedures just described. We generated data from a mixture of a Bernoulli distribution and a lognormal distribution with mean and variance of 1 using the same combinations of sample sizes and probabilities of observing a non-zero as was done for the frequency distributions. For all methods, coverage rates for 95% equal -tailed confidence intervals were calculated as:

1) The proportions of confidence intervals whose lower bounds (L) were below the true mean (γ);
2) The proportions of confidence intervals whose upper bounds (U) were above the true mean; and,
3) The proportion of confidence intervals that covered the true mean ($L < \gamma < U$).

For confidence intervals based on the asymptotic or MOVER methods, 5000 samples were simulated. For each sample, the asymptotic interval, asymptotic intervals based on either a natural log transformation or a square root transformation, or Zou et al.'s (2009) MOVER method for an interval were calculated. The coverage rates for the lower bound, the upper bound and the entire confidence interval for a 95% confidence level were calculated (Table 3).

For bootstrap confidence intervals a total of 100 samples were simulated from the population and for each sample 200 bootstrap samples were selected with replacement. For the nested bootstrap, an additional 25 leave-one-out jackknife samples were taken to obtain an estimate of the variance of Aitchison's estimator $\hat{\gamma}_{Ait}$. Coverage of the endpoints and the entire interval were assessed based on the behavior of the 100 bootstrap confidence intervals.

The results are interesting – the lower bounds of all of the confidence intervals were virtually always below the true mean (error rates < 0.025) while coverage of the upper bounds and consequently for the entire interval varied by method. The upper bounds for the bootstrap intervals based on $\hat{\gamma}_{Ait}$ and the intervals based on asymptotic arguments often failed to cover the true mean of the population. Not surprisingly, the upper bounds had better coverage with larger sample sizes and when the proportion of non-zeroes in the population was large. Since the distributions of Aitchison's estimator of the mean are skewed right, small sample sizes will not necessarily capture the full range of values likely to be in the sampling distribution of $\hat{\gamma}_{Ait}$.

Although the naïve estimators ($\hat{\gamma}_{Nai}$) are biased, the intervals based on the MOVER method were well behaved in that they had upper tail error rates not far from expected (around 0.05 vs the expected 0.025) for the upper endpoint as well as the overall confidence interval.

Taking a transformation in an attempt to stabilize variance and symmetrize the distribution of $\hat{\gamma}_{Ait}$ tended to increase somewhat the coverage rates for the asymptotic or simple bootstrap intervals but at the cost of having some cases where confidence intervals could not be calculated because the estimated mean equaled 0 (and hence the variance based on the delta method was undefined). Because the MOVER method also uses a natural log-transformation, it suffers from the same failing. The inability to obtain a confidence interval is not a rare occurrence for some populations that are of interest in fisheries research and monitoring. For example, for a sample of size N = 25 from a population with a probability of a non-zero of 0.1, the chance that all 25 observations would be 0s is 7.18%. A transformation would not be of use for those samples. Note though that for an estimated mean of 0 the asymptotic interval or bootstrap interval without a transformation are calculable but non-informative since the interval endpoints are 0 as well. This has the tendency to lower the coverage rates for all methods of confidence interval estimation when sample sizes are small and there are a high proportion of 0s in the population.

Bootstrapping based on $\hat{\gamma}_{Ait}$ for percentile confidence interval estimation did not improve substantially over the asymptotic approaches (Table 3). In fact, using bootstrapping with transformations provided intervals with slightly worse coverage than the equivalent asymptotic methods (not shown). The poor coverage of bootstrapping intervals has been demonstrated for the more general case where the variance of an estimator of a parameter is itself a function of that parameter. Efron and Tibshirani (1993, pg. 163-165) discuss the failure of such bootstrap intervals. They describe a five-step approach involving nested bootstrapping to stabilize the variance and estimate a confidence interval in this situation. It involves not only the estimation of the variance using a second bootstrap (rather than using a defined function such as $\widehat{V}_{Pen}[\widehat{\gamma}_{Ait}]$ for estimating the variance of the mean) as we did here but going further and also estimating a transformation that would further stabilize the variance. Perhaps this procedure could be used here as well. We did not consider it since the nested bootstrap-t method performed very well, having coverage rates for upper and lower bounds and for the entire confidence interval that were near 0.95. Interestingly, the coverage approaches 0.95 from above as the sample size increases. The nested bootstrap-t method is preferred over any other method considered except the MOVER approach, the best non-resampling method for confidence interval estimation of a delta-lognormal distribution. If computational aspects are of concern, there have been several papers discussing methods for reducing the computational effort (*cf*. Hinkley & Shi, 1989, Newton & Geyer, 1994).

In some instances, especially at small sample sizes, interval estimation methods provided good coverage but at the cost of very wide interval estimates. For example, the nested bootstrap method had a coverage of 0.95 at $N = 50$ and $p = 0.1$ but an expected length of 84; the MOVER method had a similar length at the same parameter and sample size combination. The expected length decreased quickly with larger sample sizes for both of the best methods.

Figure 1. Histograms of Aitchison's estimator for various sample sizes and probability of observing a non-zero. Row 1 : N = 25; row 2, 50; and, row 3, 100. Column 1: p = 0.1; column 2, 0.3; and, column 3, 0.6. In all instances, the underlying lognormal distribution of the non-zero data had a mean and variance of 1.

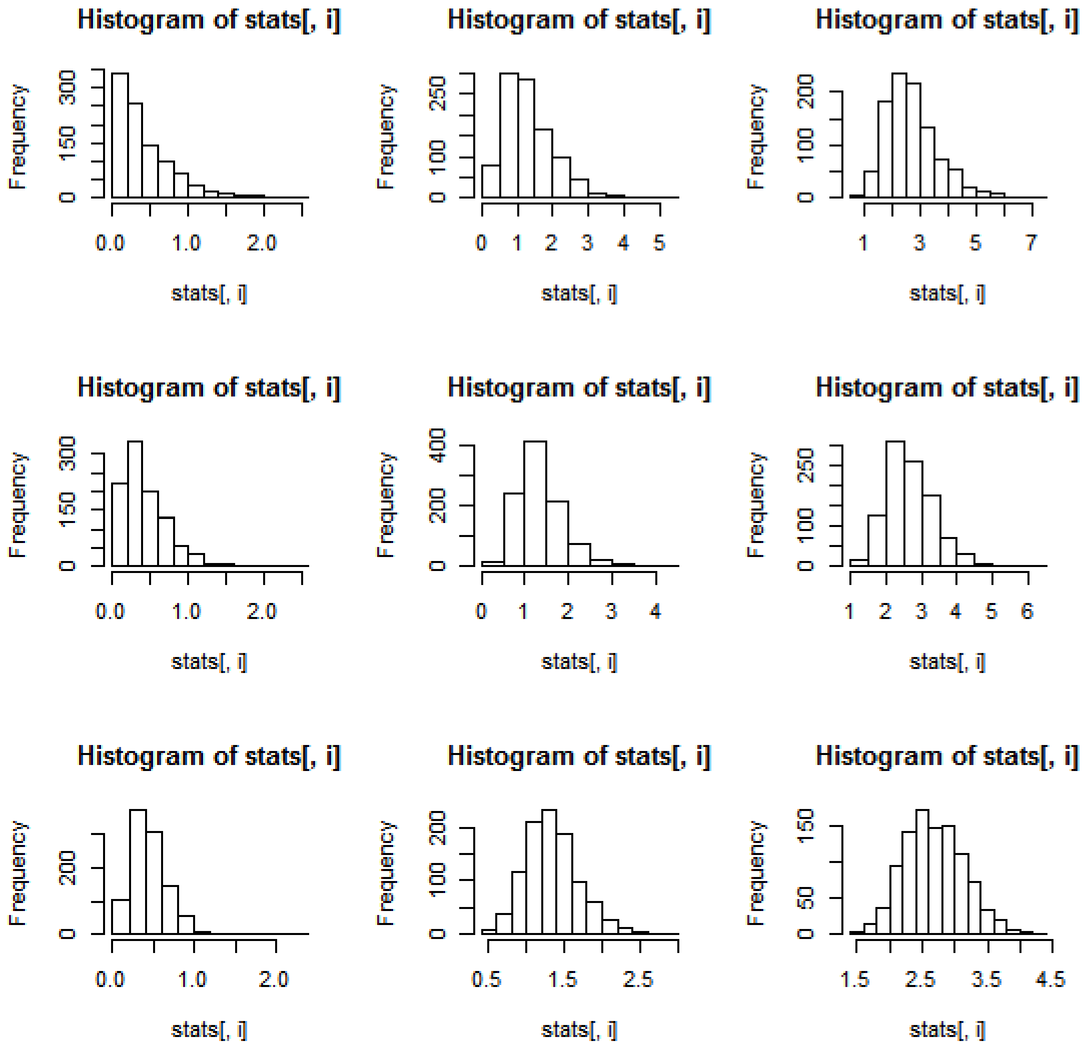


Figure 2. Histograms of Aitchison's estimator of the mean of the mixture distribution for various combinations of sample sizes and the proportion of non-zeroes. Row 1 : N = 25; row 2, 50; and, row 3, 100. Column 1: p = 0.1; column 2, 0.3; and, column 3, 0.6. In all instances, the underlying lognormal distribution of the non-zero data had a mean of 1 and variance of 4.

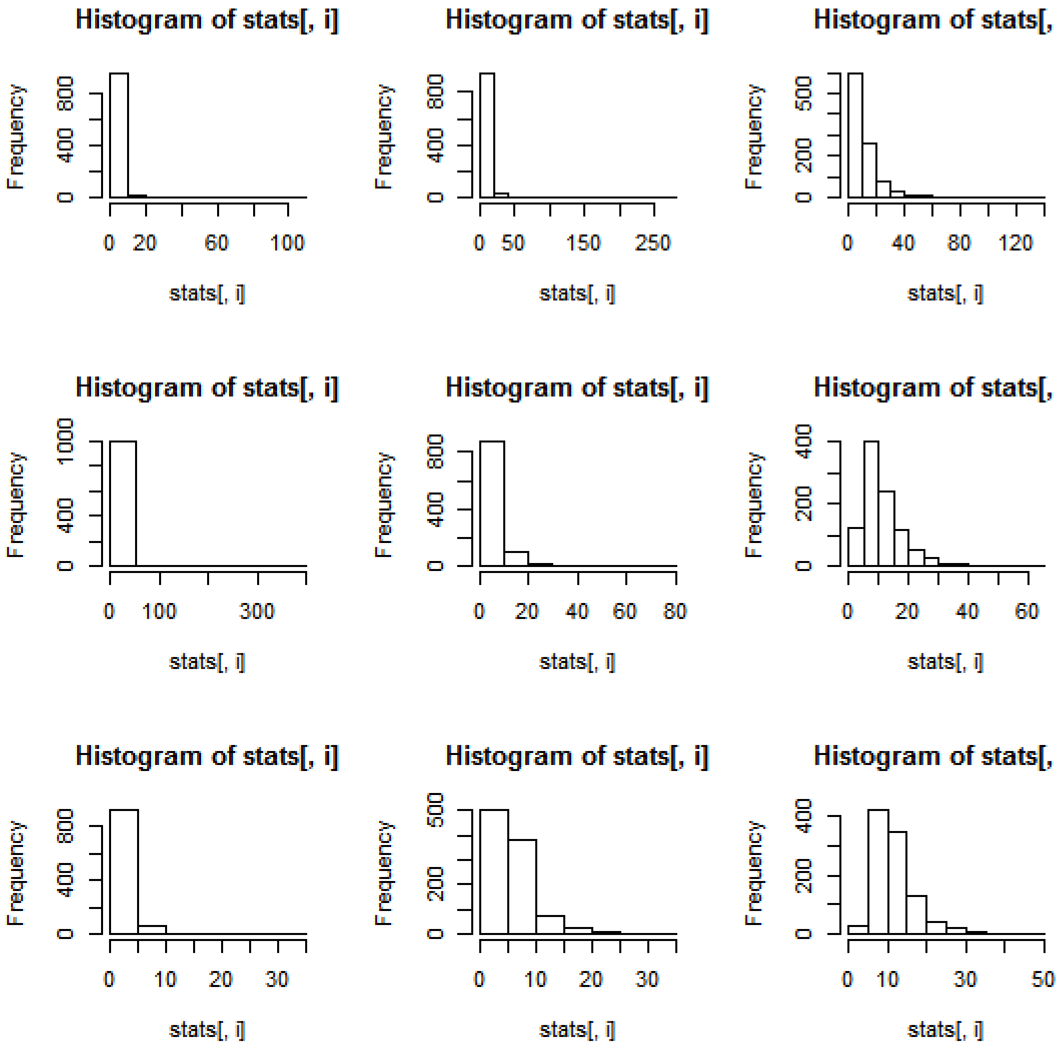
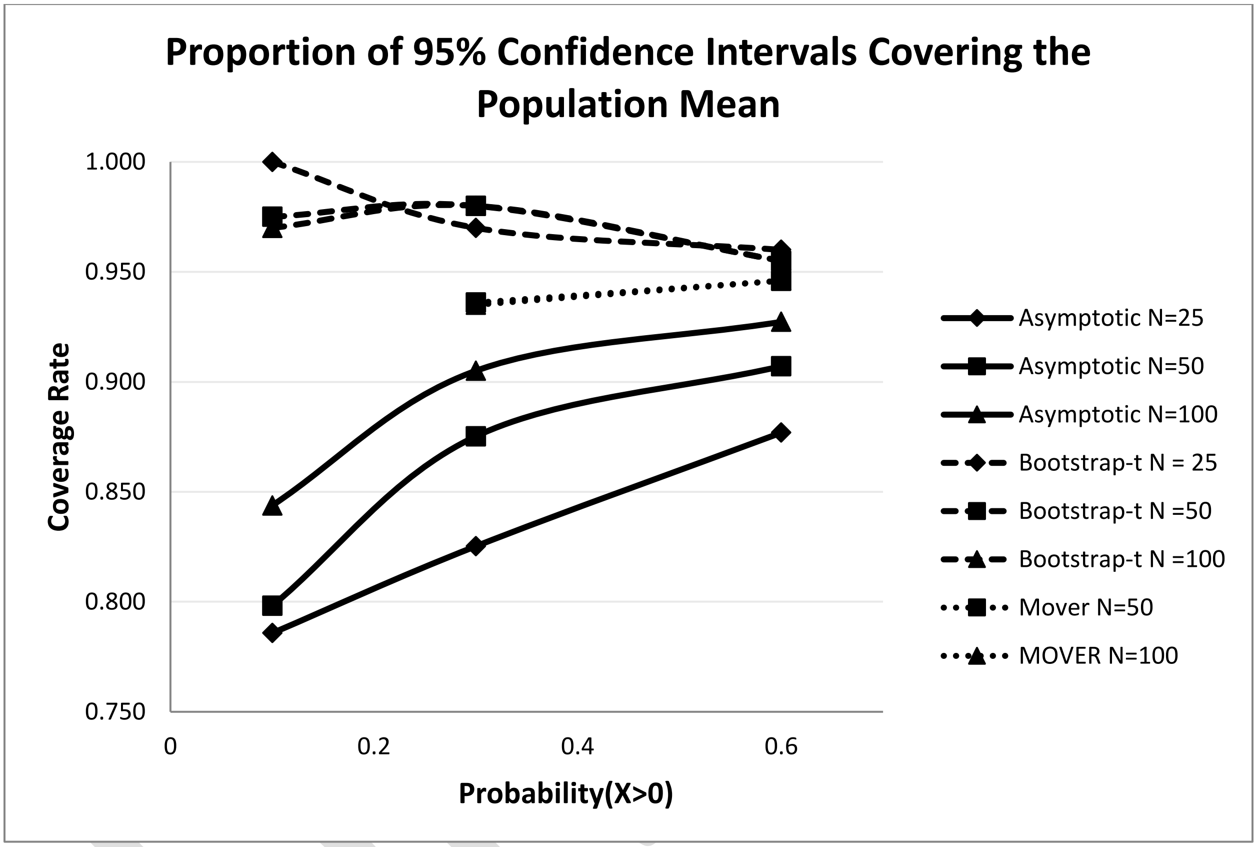


Table 3. Coverage rates for six methods (see text) for calculating a 95% confidence interval for the mean of the mixture distribution. Rates obtained from simulations from distributions with a varying probability of observing a non-zero $(1-\theta = 0.1,\ 0.3,\ \text{or } 0.6)$, three sample sizes (25, 50, 100), and non-zero observations from a lognormal distribution with mean and variance of 1. Estimation is based on Aitchison's estimator of the mean, Pennington's estimator of the variance of the estimated mean in the asymptotic methods and one of the bootstrap methods, and a bootstrap estimator of the variance for two of the bootstrap methods. LB = proportion of simulated confidence intervals whose lower bound is less than the true mean; UB = proportion of simulated confidence intervals whose upper bound is greater than the true mean; and CI = proportion of confidence intervals which cover the true mean. EL = expected length estimated as average length of simulated confidence intervals. RB = relative bias of the confidence interval. Areas highlighted in grey represent combinations of method, sample size and proportion of non-zeroes that have intervals with approximately 90% coverage or better.

Sample Size	$1-\theta$	Asymptotic					Log-transform + Asymptotic					MOVER				
		LB	UB	CI	EL	RB	LB	UB	CI	EL	RB	LB	UB	CI	EL	RB
N=25	0.2	1.000	0.678	0.678	4.01	0.192	0.974	0.824	0.803	152.50	0.083	0.972	0.845	0.845	81.22	0.070
	0.4	1.000	0.761	0.761	6.18	0.136	0.981	0.857	0.838	7.55	0.067	1.000	0.981	0.981	4.24	0.010
	0.6	1.000	0.804	0.804	7.91	0.109	0.986	0.887	0.873	9.13	0.052	1.000	0.901	0.901	2.66	0.052
N=50	0.2	1.000	0.760	0.760	3.19	0.137	0.980	0.870	0.851	13.57	0.059	1.000	0.989	0.989	5.53	0.006
	0.4	1.000	0.829	0.829	4.90	0.093	0.985	0.908	0.893	5.53	0.040	1.000	0.920	0.920	2.21	0.042
	0.6	1.000	0.865	0.865	6.02	0.073	0.987	0.919	0.906	6.54	0.036	0.999	0.932	0.931	1.66	0.035
N=100	0.2	1.000	0.831	0.831	2.50	0.092	0.984	0.905	0.888	2.83	0.042	0.999	0.918	0.917	2.25	0.042
	0.4	1.000	0.870	0.870	3.65	0.069	0.985	0.934	0.919	3.90	0.026	0.998	0.946	0.943	1.42	0.027
	0.6	1.000	0.896	0.896	4.48	0.055	0.986	0.940	0.929	4.68	0.023	0.994	0.945	0.939	1.11	0.025
N=200	0.2	1.000	0.872	0.872	1.86	0.068	0.987	0.937	0.923	1.99	0.026	0.998	0.945	0.943	1.47	0.027
	0.4	0.999	0.915	0.914	2.68	0.044	0.991	0.946	0.936	2.78	0.023	0.993	0.948	0.941	0.97	0.023
	0.6	0.997	0.929	0.925	3.25	0.036	0.986	0.951	0.936	3.32	0.018	0.989	0.951	0.941	0.77	0.020

Sample Size	$1-\theta$	Bootstrap $\hat{\gamma}$					Bootstrap $T_P = \frac{\hat{\gamma}-\hat{E}[\hat{\gamma}]}{\hat{V}_{Pen}[\hat{\gamma}]}$					Nested Bootstrap $T_B = \frac{\hat{\gamma}-\hat{E}[\hat{\gamma}]}{\hat{V}_{Boot}[\hat{\gamma}]}$				
		LB	UB	CI	EL	RB	LB	UB	CI	EL	RB	LB	UB	CI	EL	RB
N=25	0.2	1.000	0.776	0.776	4.91	0.160	0.897	0.981	0.878	3461.42	0.045	0.996	0.952	0.948	84.45	0.023
	0.4	1.000	0.826	0.826	7.90	0.115	0.876	0.988	0.864	305.26	0.060	0.997	0.936	0.933	41.66	0.031
	0.6	0.994	0.836	0.830	8.83	0.091	0.845	0.985	0.830	58.03	0.077	0.990	0.938	0.928	19.24	0.027
N=50	0.2	1.000	0.840	0.840	3.68	0.010	0.869	0.996	0.865	78.71	0.068	0.996	0.942	0.938	13.26	0.028
	0.4	0.996	0.902	0.898	5.32	0.082	0.826	0.993	0.819	24.78	0.092	0.989	0.954	0.943	9.35	0.018
	0.6	0.998	0.928	0.926	7.03	0.048	0.866	0.955	0.861	20.37	0.049	0.985	0.958	0.943	9.94	0.014
N=100	0.2	0.998	0.888	0.886	2.77	0.068	0.873	0.996	0.869	15.70	0.068	0.966	0.961	0.957	5.58	0.018
	0.4	0.996	0.926	0.922	3.85	0.044	0.860	0.999	0.859	9.02	0.075	0.983	0.965	0.948	5.27	0.009
	0.6	0.998	0.936	0.934	4.62	0.042	0.868	0.993	0.861	8.30	0.067	0.987	0.968	0.955	5.66	0.010
N=200	0.2	0.998	0.916	0.914	1.99	0.048	0.874	0.996	0.870	4.72	0.065	0.991	0.967	0.958	2.74	0.012
	0.3	0.990	0.940	0.930	2.76	0.020	0.896	0.994	0.980	4.27	0.051	0.988	0.974	0.962	3.23	0.007
	0.6	0.986	0.952	0.938	3.37	0.021	0.905	0.995	0.900	4.63	0.047	0.983	0.972	0.955	3.73	0.006

Figure 3. Coverage rates for various populations and sample sizes for three methods (asymptotic, nested bootstrap-t based on t_B, and MOVER) of constructing confidence intervals.



Estimators Based on Models Fitted with Explanatory Variables

In this section we describe estimation of the mean of the mixture distribution and its standard error when the estimators of the probability of a non-zero and the mean of the non-zero distribution are obtained via model fitting, such as general linear models or generalized additive models. The following results are based on results from general linear models. The concepts remain the same for mixed models or non-linear or additive models; only the specific details for estimation change. Alternative modeling approaches include generalized additive models (Hastie and Tibshirani, 1990; Wood, 2006, 2004; and others), LASSO regression (Tibshirani, 1996), generalized linear models (McCullogh & Nelder, 1989), classification and regression trees (Breiman, et al. 1984) and others. Regardless of the statistical method used, each of the models provide a predicted value, i.e. an estimated mean, for each observation in the data set.

Zero-Altered Model Approach

For the following, construct two random variables from the observations:

$$1)\quad X^P = \begin{cases} 1, & \text{if } X = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$2)\quad Y = \begin{cases} ., & \text{if } X = 0 \\ \ln(X), & \text{otherwise} \end{cases}$$

where ‘ . ’ indicates a missing value and X is, as before, the value of the observation. We start by modeling each of these variables on a set of explanatory variables, such as temperature or chlorophyll a concentration, using standard statistical modeling methods. For example we show results for when the indicator variable for presence, X^P, is modeled using logistic regression with a logit link. The transformed non-zero data, Y, is modeled using a general linear model under the assumption that the natural log transformed data are normally distributed.

Binomial Component

Assuming that $\{X_i^P\}_{i=1}^n$ are distributed as independent Bernoulli random variables with means $(1-\theta_i)$, where $\theta_i = \Pr(X_i^P = 0)$, a logistic regression model relating $\boldsymbol{\theta}' = \{\theta_1, \dots, \theta_n\}$ to the explanatory variables is

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \eta_i = \boldsymbol{z}_i\boldsymbol{\beta}, \qquad i = 1, \dots, n$$

or equivalently

$$\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}' = \{\beta_0, \beta_1, \dots, \beta_{(s-1)}\}$ is the column vector of model coefficients and $\boldsymbol{Z}$ be a $n \times s$ matrix with rows $\boldsymbol{z}_i = \left\{1, x_{1i}, \dots, x_{(s-1)i}\right\}, i = 1, \dots, n$. The row vector $\boldsymbol{z}_i = \left\{1, x_{1i}, \dots, x_{(s-1)i}\right\}$ contains a leading 1 to allow for an intercept term and the remaining elements are the values for the $s-1$ explanatory variables or functions of the explanatory variables such as interactions. By *a priori* setting some of the coefficients to zero, the researcher can control which set of explanatory variables are used in each of the component models.

Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of the vector of coefficients given the data (usually obtained using the Newton-Raphson method); then the fitted values on the linear scale are $\widehat{\boldsymbol{\eta}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$. The variance of $\hat{\eta}_i$ is $var(\hat{\eta}_i) = \boldsymbol{z}_i \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{z}_i{}'$ where $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = var(\widehat{\boldsymbol{\beta}})$ is the $k \times k$ variance-covariance matrix of the estimated coefficients. The usual estimator of $var(\hat{\eta}_i)$ is $\widehat{var}(\hat{\eta}_i) = \boldsymbol{z}_i \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{z}_i{}'$ where $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}$ is the estimated variance-covariance matrix obtained as the inverse of the Hessian matrix evaluated at the model coefficients estimates. The estimated values $\widehat{\boldsymbol{\eta}}$ and their standard errors $\sqrt{\widehat{var}(\hat{\eta}_i)}$ are given by most software packages. See Agresti (2002) for more detail on fitting logistic regression models.

The maximum likelihood estimator of the probability that the i^{th}observation has a zero abundance is the back-transformation

$$\hat{\theta}_i = \frac{\exp(\boldsymbol{z}_i\widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{z}_i\widehat{\boldsymbol{\beta}})} = \frac{1}{1 + \exp(-\boldsymbol{z}_i\widehat{\boldsymbol{\beta}})}.$$

The approximate variance of $\hat{\theta}_i$ is again based on the Delta Method and is given by

$$var\left(\hat{\theta}_i\right) = [\theta_i(1 - \theta_i)]^2 var(\hat{\eta}_i).$$

The variance of $\hat{\theta}_i$ is estimated by replacing the parameters with their estimates:

$$\widehat{var}\left(\hat{\theta}_i\right) = \left[\hat{\theta}_\imath\left(1 - \hat{\theta}_\imath\right)\right]^2 \widehat{var}(\hat{\eta}_i).$$

Comment: The Delta Method for approximating variances of functions of random variables relies on a first-order Taylor expansion around the mean. In order for the approximation be close to the true value it is required that the function (here the logit) be linear over the range of likely values (Cook, 2008). Hence, if the function is highly non-linear, the delta method will give a poor approximation to the estimated variance. In such cases it is recommended that the Taylor series on which the Delta method is based include additional terms.

The estimated variance of $\hat{\theta}_i$ using the plug-in approach above may provide a biased estimate of the variance. Hence, it has often been suggested that alternative approaches be used to obtain better approximations to the desired variance. Parametric bootstrapping or Monte Carlo methods are two that are often mentioned (*cf*. Davison & Hinkley, 1997).

Lognormal Component

Here we assume that the set $\{\ln(X_i)\}_{i=1}^{m}$, where $m(< n)$ is the number of non-zero observations, are independently lognormally distributed with means and variances

$$\begin{array}{rl} E[\ln(X_i)\,; X_i > 0] = \mu_i = \boldsymbol{z_i\xi} & \\ V[\ln(X_i); X_i > 0] = \sigma_i^2 = \sigma^2, & \end{array} \qquad i = 1, \dots, m$$

where $\boldsymbol{\xi}$ is the column vector of model coefficients and $\boldsymbol{Z_s}$ is the $m \times k$ matrix of explanatory variable values for the m available observations and $k - 1$ explanatory variables. The $\boldsymbol{Z_s}$ matrix could contain the same explanatory variables, a subset, or a completely different set of potential predictors from those used in the fitting of the binomial component of the model. The assumption of homoscedastic error variances can be relaxed to allow for unequal variance as well.

If this model is fitted to the m available observations, we obtain the maximum likelihood estimates for the coefficients $\hat{\boldsymbol{\xi}} = (\boldsymbol{Z_s}'\boldsymbol{Z_s})^{-\mathbf{1}}\boldsymbol{Z_s}'\boldsymbol{Y}$ where $\boldsymbol{Y}' = [\ln(X_1)\,, \dots, \ln(X_m)]$ are the observed values. The variance-covariance matrix for $\hat{\boldsymbol{\xi}}$ is $var(\hat{\boldsymbol{\xi}}) = \sigma^2(\boldsymbol{Z_s}'\boldsymbol{Z_s})^{-\mathbf{1}}$. The estimator for σ^2 is the model MSE given by $MSE = \hat{\sigma}^2 = \frac{1}{(m-k)}\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{Z_s}(\boldsymbol{Z_s}'\boldsymbol{Z_s})^{-\mathbf{1}}\boldsymbol{Z_s}')\boldsymbol{Y}$.

Now, of interest is the estimator of the non-zeroes given presence back on the original scale (not the log) scale and additionally an estimate of the variance of the back-transformed estimates. We know that, in general, for a random variable X that is lognormally distributed such that $\ln(X) \sim N(\mu, \tau^2)$, the mean and variance of X are

$$E(X; X > 0) = \alpha = \exp\left(\mu + \frac{\tau^2}{2}\right)$$

$$V(X; X > 0) = \exp(2(\mu + \tau^2)) - \exp(2\mu + \tau^2)$$

(Casella & Berger, 2002). Goldberger (1968) derived an unbiased estimator of α when the mean $\mu_i = \boldsymbol{z}_i\boldsymbol{\xi}$ is a function of covariates:

$$\hat{E}(X_i; X_i > 0) = \hat{\alpha}_i = \exp(\boldsymbol{z}_i\hat{\boldsymbol{\xi}})\, g_{(m-k)}\left[\left[\frac{m-k+1}{2(m-k)}\right]\hat{\sigma}^2(1 - \boldsymbol{z}_{s,i}'(\boldsymbol{Z_s}'\boldsymbol{Z_s})^{-\mathbf{1}}\boldsymbol{z}_{s,i})\right]$$

$$= \exp(\boldsymbol{z}_i\hat{\boldsymbol{\xi}})\, g_{(m-k)}\left[\left[\frac{m-k+1}{2(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]$$

where $\hat{\sigma}^2_{\boldsymbol{z}_i\hat{\boldsymbol{\xi}}} = \hat{\sigma}^2\boldsymbol{z}_{s,i}'(\boldsymbol{Z_s}'\boldsymbol{Z_s})^{-\mathbf{1}}\boldsymbol{z}_{s,i}$ is the estimated variance of $\hat{\mu}_i = \boldsymbol{z}_{s,i}\hat{\boldsymbol{\beta}}$ and $g_{(m-k)}(t)$ is the generalized hypergeometric series first described by Finney (1941).

Comment: The quantity $\sigma^2(1 - \mathbf{z}_{s,i}{}'(\mathbf{Z}_s{}'\mathbf{Z}_s)^{-1}\mathbf{z}_{s,i})$ is the variance of the residual associated with the i^{th} observation.

For estimating variance, Bradu and Mundlak (1970) have shown that an unbiased estimator for functions $\exp(\omega\mathbf{z}_{s,i}\boldsymbol{\xi} + \varphi\sigma^2)$, where ω and φ are non-zero constants, is

$$\widehat{\exp}\left(\omega\mathbf{z}_{s,i}\boldsymbol{\xi} + \varphi\sigma^2\right) = \exp\left(\omega\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}\right) g_q\left[\frac{q+1}{q}\left(\varphi\hat{\sigma}^2 - \frac{1}{2}\omega^2\hat{\sigma}^2_{\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]$$

where q is the number of degrees of freedom associated with $\hat{\sigma}^2$.

Using this, they obtained an estimator for

$$V(X_i; X_i > 0) = \exp\left(2(\mathbf{z}_{s,i}\boldsymbol{\xi} + \sigma^2)\right) - \exp(2\mathbf{z}_{s,i}\boldsymbol{\xi} + \sigma^2)$$

which is given by

$$\hat{V}(X_i; X_i > 0) = \exp\left(2\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}\right)\left\{g_{(m-k)}\left[\left[\frac{2(m-k+1)}{(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]\right.$$
$$\left. - g_{(m-k)}\left[\left[\frac{(m-k+1)}{(m-k)}\right]\left(\hat{\sigma}^2 - 2\hat{\sigma}^2_{\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]\right\}.$$

Bradu & Mundlak (1970) extended Goldberger's results and derived an estimate of the variance of Goldberger's estimator of $E(X; X > 0)$:

$$\widehat{Var}(\hat{\alpha}_i) =$$
$$\exp\left(2\mathbf{z}_i\hat{\boldsymbol{\xi}}\right)\left\{g^2_{(m-k)}\left[\left[\frac{m-k+1}{2(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{\mathbf{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right] - g_{(m-k)}\left[\left[\frac{m-k+1}{(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{z_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]\right\}.$$

Shono (2008) recommended a biased procedure where the mean $E(X; X > 0)$ is estimated using the predicted median:

$$\hat{E}(\boldsymbol{X}; \boldsymbol{X} > 0) = \exp(\hat{\boldsymbol{\mu}}) = \exp\left(\boldsymbol{Z}_s\hat{\boldsymbol{\xi}}\right).$$

This estimator is for the median and so will be biased low since for right-skewed distributions the mean > median.

Combined Zero-Altered Model

Predictions on the original data scale are obtained as a product of all of the observed data (zeroes and non-zeroes). This is accomplished in a manner similar to that used by the Aitchison-

Pennington estimator of the mean except that the mean of the normal distribution of the log-transformed data is obtained from the fitted model as described above and the estimated probability of a zero value being observed is obtained from the logistic regression model fitted to the presence/absence data.

The first step is to obtain the estimated values for θ_i for all n observations. These are estimated as part of the fitting of the logistic regression model to all n observations. Next, the fitting of the general linear model to the m observations of non-zero data provides estimated coefficients for the linear regression model that can be used to obtain predicted values not only for the m observations but also for the remaining $n - m$ observations. One caveat to this method is that the values of the covariates for the $n - m$ observations must meet the following conditions: 1) the values of the elements of the vectors $\{\boldsymbol{z}_{\boldsymbol{s},\boldsymbol{i}}, i = m + 1, \dots, n\}$ of the zero-values for Y fall within the $(s - 1) -$ dimensional space of the vectors $\{\boldsymbol{z}_{\boldsymbol{s},\boldsymbol{i}}, i = 1, \dots, m\}$ used in the model fitting; and 2) the pattern of multicollinearity among the covariates, if it exists, must be the same among the $\{\boldsymbol{z}_{\boldsymbol{s},\boldsymbol{i}}, i = m + 1, \dots, n\}$ as among the $\{\boldsymbol{z}_{\boldsymbol{s},\boldsymbol{i}}, i = 1, \dots, m\}$.

The predicted values for $X_i, i = 1, \dots, n$, which are also estimators of the means of the mixture distribution, are given by

$$\begin{aligned}\widehat{E_M[X_\imath]} = \hat{\gamma}_{i,M} &= \left(1 - \hat{\theta}_i\right)\hat{\alpha}_i \\ &= \left(1 - \hat{\theta}_i\right) \exp\left(\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}\right) g_{(m-k)}\left[\left[\frac{m-k+1}{2(m-k)}\right]\hat{\sigma}^2\left(1 - \boldsymbol{z}_{s,i}{}'\left(\boldsymbol{Z_s}{}'\boldsymbol{Z_s}\right)^{-\mathbf{1}}\boldsymbol{z}_{s,i}\right)\right].\end{aligned}$$

The variance of $\hat{\gamma}_{i,M}$ is obtained using

$$Var\left[\hat{\gamma}_{i,M}\right] = Var\left[\left(1 - \hat{\theta}_i\right)\hat{\alpha}_i\right] = \theta_i(1 - \theta_i)\{E[\hat{\alpha}_i]\}^2 + (1 - \theta_i)Var[\hat{\alpha}_i]$$

(Goodman, 1960) which can be estimated by replacing the set of parameters

$$\{\theta_i, \{E[\hat{\alpha}_i]\}^2, Var[\hat{\alpha}_i]\}$$

with their point estimators

$$\{\hat{\theta}_i,\ \exp\left(2\boldsymbol{z}_{s,i}\boldsymbol{\xi}\right) g_{(m-k)}\left[\left[\frac{(m-k+1)}{(m-k)}\right]\left(\hat{\sigma}^2 - 2\hat{\sigma}^2_{\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right], \widehat{Var}(\hat{\alpha}_i)\}$$

described above. Hence, the estimated variance of $\widehat{E_M[X_\imath]} = \hat{\gamma}_{i,M}$ is

$$\widehat{Var}[\hat{\gamma}_{i,M}] = \hat{\theta}_i(1-\hat{\theta}_i)\exp(2\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}})\,g_{(m-k)}\left[\left[\frac{(m-k+1)}{(m-k)}\right]\left(\hat{\sigma}^2 - 2\hat{\sigma}^2_{\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]$$
$$+ (1-\hat{\theta}_i)\exp(2\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}})\left\{g^2_{(m-k)}\left[\left[\frac{m-k+1}{2(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]\right.$$
$$\left. - g_{(m-k)}\left[\left[\frac{m-k+1}{(m-k)}\right]\left(\hat{\sigma}^2 - \hat{\sigma}^2_{\boldsymbol{z}_{s,i}\hat{\boldsymbol{\xi}}}\right)\right]\right\}.$$

Failures of Assumptions

There are three assumptions that have been used so far in our study:

1) that the two estimators $\hat{\theta} = \widehat{Prob}(X = 0)$ and $\hat{\alpha} = \hat{E}[X; X > 0]$ are independent random variables;
2) that the Bernoulli and Log-Normal distributions are in fact the distributions of the respective derived random variables $X^P = I[X = 0]$ and $Y = \ln(X)$; and,
3) that the sampling design is a simple random sample so that the observations are independent.

Non-Independence of the Estimators

Before discussing the effect of non-independence of the two estimators of the mean parameters of the two components of the delta-distribution mixture on estimation, we review some definitions of random variables, covariance and correlation and their influence on the expectation and variance of products of random variables. As we will see, the determination of non-independence requires a clear understanding of these definitions. Once established, we discuss conditions under which we might expect the two estimators to be correlated and when they are not correlated.

Definition: The set S of all possible outcomes of a particular experiment is called the $sample\ space$ of the experiment.

Definition: Given a random experiment (an experiment in which the outcome is not known *a priori*) with sample space S, a $random\ variable$ is a function X that maps the original sample space into a new sample space, usually the set of real numbers.

For example, suppose we randomly sample 100 items from some population (a random experiment) and record for each item whether its length exceeds 50 cm. The sample space of this experiment is the 2^{100} possible ordered lists of “Yes” or “No” measurements, each list of length 100. This is a random experiment since we do not know which ordered list will be observed until the experiment has been performed and, further, repetitions of the experiment will yield different ordered lists. The number of the 100 items that exceed the cutoff is a random variable that maps from the sample space of ordered lists of the 100 indicators {Yes, No} to a new sample space, S_X say, with integer values ranging between 0 and 100. In other words, X is a random variable that in this case summarizes information in a random outcome from an experiment.

A random variable has a probability distribution

$$\Pr(X = x)\,, \text{for } x \in S_X$$

for discrete random variables or, for continuous random variables, a probability density function

$$f(x), x \in S_X \text{ such that } \int_{x \in S_X} f(x)dx = 1.$$

For example, in our earlier experiment with random sampling of items, the outcome X has a probability distribution given by the Binomial probability distribution with parameters $n = 100$ and $\theta = \Pr(\text{item's length > 50 cm})$.

If two random variables are used to map the outcome from an experiment they have a joint distribution,

$$\Pr(X_1 = x_1, X_2 = x_2)\,, \text{for } x_1 \in S_{X_1} \text{ and } x_2 \in S_{X_2}$$

for discrete random variables and for continuous random variables, a probability density function

$$f(x_1, x_2), x_1 \in S_{X_1} \text{ and } x_2 \in S_{X_2} \text{ such that } \int\int f(x_1, x_2)dx_1dx_2 = 1$$

that describes the probability or likelihood of observing a particular outcome or value. For example, in our earlier experiment define X_1 to be the number of items whose lengths exceed 50 cm and let X_2 be the length of the longest run of sequential “Yes” measurements. Then, $\Pr(X_1 = 27, X_2 = 15)$ is the probability that exactly 27 “Yes”es were observed and that the longest run of sequential “Yes”es in the string of 100 measurements was 15 in a row.

In a joint distribution, we have the parameters

$$\mathrm{E}[X_1] = \mu_1, \mathrm{V}[X_1] = \sigma_1^2,$$

$$\mathrm{E}[X_2] = \mu_2, \mathrm{V}[X_2] = \sigma_2^2,$$

and

$$Cov[X_1, X_2] = \sigma_{12}.$$

Definition: Let A and B be two random variables with means μ_A and μ_B and variances σ_A^2 and σ_B^2, respectively. The covariance of A and B is

$$Cov(A, B) = E[(A - \mu_A)(B - \mu_B)] = E[AB] - \mu_A \mu_B.$$

The correlation is

$$Corr(A, B) = \frac{Cov(A, B)}{\sqrt{\sigma_A^2 \sigma_B^2}}.$$

Two variables are said to be uncorrelated if $Cov(A, B) = 0$, i.e. when $E[AB] = E[A]E[B] = \mu_A \mu_B$. Note that covariance refers to the relationship of two random variables with a joint probability distribution and is estimable only of we have several independent observations (x_1, x_2) from the joint distribution of the two random variables (X_1, X_2).

If a simple random sample of n pairs of observations $\{a_i, b_i\}_{i=1}^{n}$ is selected from the joint distribution of A and B, then the UMVUE of $Cov(A, B)$ is given by

$$\widehat{Cov}(A, B) = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \hat{\mu}_A)(b_i - \hat{\mu}_B)$$

(Lehmann, 1983). The correlation is estimated using

$$\widehat{Corr}(A, B) = \frac{\widehat{Cov}(A, B)}{\sqrt{\hat{\sigma}_A^2 \hat{\sigma}_B^2}}.$$

A common problem with using an estimator that is a product of two correlated estimators is that often we have data from only a single sampling event, i.e. for example, for the mixture distribution we are reviewing in this paper, we have only one value for $\hat{\theta}^c = \left(1 - \hat{\theta}\right)$ and one for $\hat{\alpha} = \hat{E}[X; X > 0]$. As a result, $Cov(\hat{\theta}^c, \hat{\alpha})$ is not estimable unless we have additional information independent of the sample. One approach that has been recommended for ongoing monitoring programs (Lo, *et al.* 1992), where there are repeated measurements such as annual surveys, is to estimate the covariance by estimating annual values for Z and α and assuming that the annual estimates, $\left\{\hat{\theta}_t^c, \hat{\theta}_{t+1}^c, \dots, \hat{\theta}_{t+s}^c;\ \hat{\alpha}_t, \hat{\alpha}_{t+1}, \dots, \hat{\alpha}_{t+s}\right\}$, are random selections from the ***same joint probability distribution*** [emphasis deliberate]. This requirement implies that the annual means and variances of the populations for which the estimates are desired are identically the same which is not the case in almost any managed fisheries. That is, the mean CPUE varies from year to year and so estimates of mean CPUE in one year may or may not be an estimate of the mean CPUE in a different year. If the annual estimates are replicates estimates of the SAME population (i.e. no changes over time), the covariance can then be estimated by

$$\widehat{Cov}\left(\hat{\theta}^c, \hat{\alpha}\right) = \frac{1}{s-1}\sum_{i=0}^{s}(\hat{\theta}^c_{t+i} - \bar{\bar{\theta}}^c)(\hat{\alpha}_{t+i} - \bar{\bar{\alpha}})$$

where $\bar{\bar{w}} = \frac{1}{s}\sum_{i=0}^{s} w_{t+i}$ is the mean of the annual estimates.

The question of interest here is whether the covariance $Cov\left(\hat{\theta}^c, \hat{\alpha}\right)$ is non-zero and hence requiring estimation at all. In the next section we discuss whether in fact the two estimators have non-zero covariance. Following that, if they are correlated, we show the impact of that on estimation of the delta lognormal mean and variance.

Are the Estimators for the Delta-Lognormal Mixture Parameters Dependent?

In the previous sections, we assumed that the two estimators, $\hat{\theta}$ and $\hat{E}[X; X > 0]$ are uncorrelated, i.e. $Cov\left(\hat{\theta}, \hat{E}[X; X > 0]\right) = 0$. In this section, we discuss the joint distribution of the two estimators, $\hat{\theta}$ and $\hat{E}[X; X > 0]$ and their covariance. A natural question is whether the two estimators co-vary. This dependence has influence on the biasedness and variance of the product of the estimators (see next section).

There are two situations in which one might think there is a non-zero covariance between the two estimators:

1) The parameters of the two distributions are functionally related, i.e. one is a function of the other; and,
2) The parameters of the two distributions are both functions of a set of covariates such as environmental predictors.

Functionally Related Parameters of the Component Distributions

Parameters Constant in Space and Time

The first instance has been used to argue that the estimators of presence and abundance are correlated, for example arguing that if the probability of a non-zero is high, then the mean of the non-zero observations is also likely to be high. It is possible for the two parameters, θ and μ, to have a restricted parameter space while having their estimators not covary. This relationship between the parameters has been confused in the literature (Lo et al. 1992; Ingram et al. 2010; Walter & Ortiz, 2011) with the question of whether estimators based on random samples for these parameters are correlated. In fact, we show they are not.

To reiterate, the fact that one parameter can be shown to be a function of the other does not in itself cause the estimators to have a non-zero covariance. To see this, suppose that the probability of a non-zero is $\theta^c = (1-\theta)$ and that the mean of the non-zero distribution is $E[X; X > 0] = k\theta^c$ where $k > 0$ and $0 < \theta < 1$. A simple random sample from the mixture distribution would yield an estimate $\hat{\theta}^c = m/n$ for the probability of observing a non-zero and an estimate for the mean of the non-zero distribution $\hat{E}[X; X > 0] = \hat{\alpha}$ that is not an explicit function of $\hat{\theta}^c$, e.g. the sample mean of the non-zeroes or the Finney (1941) estimator for the mean of a lognormal distribution.

The covariance between these two estimators is

$$Cov(\hat{\alpha}, \hat{\theta}^c) = E\left[\hat{\alpha}\hat{\theta}^c\right] - \alpha\theta^c = E_{\hat{\theta}^c}\left[E\left[\hat{\alpha}\middle|\hat{\theta}^c\right]\right] - \alpha\theta^c = E\left[\alpha\hat{\theta}^c\right] - \alpha\theta^c = 0$$

where $E[A|B]$ is the conditional expectation of the random variable A given the value of the random variable B. An important point here is that we $\hat{\alpha} = f(\{X_1, \dots, X_m\}, n, m)$ is a function of the random sample of non-zero values $\{X_1, \dots, X_m\}$ and the sample sizes n and m. Hence, $\hat{\alpha}$ is a function of $\hat{\theta}^c$ only through the number of non-zero observations, m, and regardless of the value of m, $\hat{\alpha}$ is conditionally (and unconditionally) unbiased for μ, i.e.

$$E\left[\hat{\alpha}\middle|\hat{\theta}^c\right] = E[f(\{X_1, \dots, X_m\}, n, m)|m, n] = E[\hat{\alpha}] = \alpha,$$

the mean of the non-zeroes. The conditional expectation $E\left[\hat{\alpha}\middle|\hat{\theta}^c\right]$ would be different than the unconditional expectation ONLY if the mean of the observed values of the sample $\{X_1, \dots, X_m\}$ depends on not on θ^c but on the observed value of θ^c, $\hat{\theta}^c$, which cannot occur.

Now, if the estimator of the mean $\hat{E}[X; X > 0] = \hat{\alpha}$ is a function of the estimator $\hat{\theta}^c$, as could happen if, for example, one could use $\hat{\alpha} = k\hat{\theta}^c$, where k is known, then the two estimators would be strongly correlated. But the common estimators in use are not of this form.

Note that the independence of the estimators is not to say that the two parameters are functionally unrelated. What it does show is that for a given sampling effort where it is presumed that all observations are random selections from the same distribution, i.e. the parameters (θ, μ, σ^2) of the mixture distribution are constants, the two estimators are not correlated. This implies that for a single sampling event (e.g. a single cruise) or even for repeated sampling events (e.g. annual monitoring cruises) from the same mixture distribution (fixed parameter values over space and time), the estimators are uncorrelated and so, the developments in earlier sections of this monograph for independent estimators are valid. The covariance is zero. We show that the perceived covariance observed in Lo *et al.* 1992 or argued as existing in Walter & Ortiz (2011) is due to the fact that the annual estimates are for parameters describing the means and variances of mixture distributions that are varying over time.

Parameters Varying over Space or Time

Suppose instead, that the parameter values are functionally related so that at any given point in time they are two constants but these constants vary over time, i.e. (θ_t^c, μ_t) are different for different times t. For example suppose they vary among years, e.g. the mean abundance and the probability of observing a non-zero are both increasing over time as might occur for a recovering fishery, but are constant within a year (or at least during the sampling season). When we use a random sample $\{X_1, \dots, X_n\}$ in any given year to calculate our estimates $\{\hat{\theta}_t^c, \hat{\alpha}_t\}$ we are in fact assuming that the X_i are random samples from the same distribution with a single set of parameter values, i.e. the X_i are identically and independently distributed where the parameters do not change over the time period or spatial coverage of the sampling.

Let the estimator for the annual probability of a non-zero be $\hat{\theta}_t^c = m_t/n_t$ and the estimator of the mean of the non-zero distribution be the annual Finney (1941) estimator, $\hat{\alpha}_t$. Then after s years of monitoring the set $\left\{\hat{\theta}_t^c, \hat{\theta}_{t+1}^c, \dots, \hat{\theta}_{t+s}^c;\ \hat{\alpha}_t, \hat{\alpha}_{t+1}, \dots, \hat{\alpha}_{t+s}\right\}$ is available. Here each set of yearly estimators $(\hat{\theta}_t^c, \hat{\alpha}_t)$ is an unbiased, uncorrelated set of estimators of the parameters (θ_t^c, μ_t) within the t^{th} year. It has been recommended (Lo *et al*. 1992) that these annual estimates be used to calculate the covariance between $(\hat{\theta}_t^c, \hat{\alpha}_t)$ for any given year. But this approach assumes that the annual estimates are replicates estimating the same parameters over the entire time series. If they are replicates, the covariance is zero as has been already shown, and if they are not, we show that any estimate of covariance is capturing the relationship of the parameters not a covariance of the estimators.

It is virtually impossible that the annual estimates are from the same distribution because of the unlikely event that the parameters of the delta-lognormal distribution, (θ, μ, σ^2), are constant over time. Hence, the estimates are NOT replicates that could be used to estimate a covariance of two random variables from a joint distribution. The sets $(\hat{\theta}_t^c, \hat{\alpha}_t)$ are from different joint distributions and so any estimated covariance as described in Lo *et al.* (1992) is capturing the relationship of the parameters not the covariance of the random variables. That is not the definition of covariance of random variables and in fact has no effect on the estimation of the parameters for any given year. We shall see that if we consider a slightly different situation where the two parameters are functionally related through their relationship to exogeneous variables we obtain the same result.

Parameters as Functions of Same Exogenous Variables

We now consider the case where the two parameters are functions of exogenous variables so that the two estimators are functions of those variables. There are two separate issues here. Recall that models using the exogenous variables to predict/estimate the mean of the distribution of the response variables assume that 1) the exogenous covariates are fixed and known, and 2) that for each set of exogenous covariate values, the response variable has a conditional distribution with a mean that is a function of those exogenous variables.

For the Bernoulli distribution of the presence/absence indicator, the typical logistic regression model has that $p_i = \exp(\boldsymbol{z_i\delta})/(1 + \exp(\boldsymbol{z_i\delta}))$ where $\boldsymbol{z_i}$ is the row vector of values for the s exogenous variables for the i^{th} level of $\boldsymbol{z_i}$. For the lognormal distribution for the non-zero random variable, the generalized linear model is based on the assumption that $E[Y_i] = E[\log(X_i)] = \mu_i = \boldsymbol{z_i\beta}$. The mean of X_i is given by $\alpha_i = \exp\left(\boldsymbol{z_i\beta} + \frac{\sigma_y^2}{2}\right)$ where $\sigma_y^2 = var(Y_i)$. Note the two parameters, p_i and μ_i, are functionally related, in this case through a set of common covariates. This implies that, like our earlier case, for a fixed value of $\boldsymbol{z_i}$ the covariance between the estimators $\hat{p}_i = \exp(\boldsymbol{z_i\hat{\delta}})/(1 + \exp(\boldsymbol{z_i\hat{\delta}}))$ and $\hat{\mu}_i = \boldsymbol{z_i\hat{\beta}}$ is 0**.** If the two models are fitted separately, the covariances between the elements of $\boldsymbol{\hat{\delta}}$ and elements of $\boldsymbol{\hat{\beta}}$**,** i.e. $Cov(\hat{\delta}_i, \hat{\beta}_j), i = 1, \dots, m, j \neq i$, are zero as well. Note that this does not refer to the covariance of coefficients within a single model, i.e. $Cov(\hat{\delta}_i, \hat{\delta}_j)$ or $Cov(\hat{\beta}_i, \hat{\beta}_j)$, which are not zero (Neter, *et al.* 1990).

We ran a small simulation to show the effect of the use of the same set of predictor variables to estimate the mean non-zero value and the probability of a non-zero value. We first constructed a sequence of values for a single predictor variable, z, ranging from -3 to +3 in increments of 0.05 (N=121). The vector of probabilities of observing a 1 for the N samples was set to $p_i = \exp(2z_i)/(1 + \exp(2z_i))$. A random value of 0 (absence) or 1 (presence) was generated from each Bernoulli distribution, from a Bernoulli with parameter p_i. This gave us a sample of N indicator variables, one for each value of $\boldsymbol{z}$, for presence or absence. A generalized linear model was fit to these N random values with predictor $\boldsymbol{z}_i = \{1, z_i\}$ and the predicted values $\boldsymbol{\hat{p}} = \exp(\boldsymbol{Z\hat{\delta}})/(1 + \exp(\boldsymbol{Z\hat{\delta}}))$ obtained. Here $\boldsymbol{\delta} = \{\delta_0, \delta_1\}$ where δ_0 is the intercept and δ_1 is the slope.

Let m be the random variable indicating the number of indicator values that equaled 1. For the non-zero component of the mixture, the vector of means for the N samples was set to $\boldsymbol{\mu} = 5\boldsymbol{Z}$. For each element of $\boldsymbol{\mu}$**,** a random value of Y was generated from a Normal distribution with mean μ_i and variance = 4. A linear model was fit using the subset of m Y–values associated with the samples where the Bernoulli random variables equaled 1 and the equivalent subset of

m $\boldsymbol{Z}$ $-$values, $\boldsymbol{Z_1}$. Here $\boldsymbol{\beta} = \{\beta_0, \beta_1\}$ where β_0 is the intercept and β_1 is the slope, so this model also allowed for estimation of an intercept. The fitted model was used to obtain predicted values $\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ for all N observations**.**

The covariance and correlation between $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{Y}}$ were then calculated using the N estimate pairs. The entire procedure (random value generation, model fitting, coefficient estimation, and covariance estimation) was repeated 25,000 times, each time storing the estimated coefficients of the two model fits, the estimated covariance and estimated correlation between the two sets of predicted values.

We start by looking at the variance - covariance matrix of the two sets of estimated coefficients $Cov(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\beta}})$, specifically the covariance between $\hat{\delta}_1$ and $\hat{\beta}_1$ which is virtually 0 (-0.001) indicating that the two estimators are independent conditional on the value of $\boldsymbol{Z}$.

The covariance between $\boldsymbol{p}$ and $\boldsymbol{\mu}$ is 3.467 (the correlation = 0.964), the same as the average covariances and correlations calculated for estimators $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{Y}}$ from the 25,000 simulations of the model fits. This supports our earlier argument that the estimate of the covariance is NOT capturing the covariance between two random variables from a joint probability distribution but is instead capturing the correlation between the parameters as they vary over the 2-D parameter space. Another way of looking at this is to assume that the values of the exogeneous variables are fixed at a single set $\boldsymbol{z_i}$ of values. Then the estimates of the means of the random variables X_i and $[X_i = 0]$, based on repeated observations from the same delta distribution, are independent as long as estimations of $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$ were done independently.

Failure of the Assumption of Lognormality of the Non-zero Observations

Another question concerning use of the Aitchison/Pennington approach for a mixture distribution is the robustness of the estimation procedure when the assumption of lognormality for the non-zero data fails. In a separate study, Christman (unpubl. ms.) used simulation to show the effect of the mis-classification of the distribution of the non-zero data. We reproduce partial results from that study here (Tables 4 and 5).

A total of 25,000 simulations were run in which data were generated from either a lognormal or an exponential distribution. The distributions were chosen so that on the log-scale, $Y = log(X)$ is either normal or exponential with a mean and variance of 1. The Aitchison/Pennington estimators and the sample mean and variance were used to estimate the means and variances of the mixture distributions on the original scale (i.e. of X, not $log(X)$), and coverage of asymptotic confidence intervals for the true means of the mixture distributions. In this study, the total sample size was set to 125, and the number of zeroes in the sample fixed at some proportion, either 0.2, 0.6, or 0.8. Hence, m the number of non-zeroes in the sample was constant at 100, 50 or 25 for every simulated sample. This removes a source of variability in the estimators and so allows us to focus on the effect of failure of the assumption of the distribution of the non-zero data.

Two results are immediate: the delta-lognormal approach severely underestimates the true mean of the mixture distribution when the $log(X)$ data are actually exponentially distributed, not log-normal, and as a consequence, the large sample CIs for the true mean $(\hat{\mu} \pm 1.96SEM)$ have actual coverage much below the nominal coverage. The sample mean and estimated SEM were much better behaved than the Aitchison/Pennington estimators when the assumption failed.

Table 4. The means of the mixture distributions used in the simulation study.

Distribution	Proportion of 1s	Mean of Mixture Dist'n
Lognormal	0.8	3.58
Lognormal	0.4	1.79
Lognormal	0.2	0.90
Exponential	0.8	11.69
Exponential	0.4	5.84
Exponential	0.2	2.92

In this study the delta-lognormal estimator was negatively biased for the true mean of the population when the distribution of $log(X)$ is exponential. Other authors have found the opposite. For example, the behavior of the estimators of $\exp\left(\mu + \frac{\sigma^2}{2}\right)$ when the non-zero data are not lognormally distributed was investigated by Myers & Pepin (1990). They started with a sample from a lognormal distribution and contaminated that sample using random selections from either a Weibull or Gamma distribution. They varied the contamination so as to impact the skew as well as the CV of the resulting distribution. They used simulations to investigate the bias and efficiency of the BLUE. They found that for fixed CV the bias increased significantly even when the amount of contamination was not too large (~20% of the data from a different distribution). Syrjala (2000), using data from several field collections, also found that the delta-lognormal estimator was sensitive to small deviations from a lognormal distribution of the non-zero data. Both authors found that the bias was positive, i.e. the mean was too high relative to the true mean, unlike our study in which the mean was biased low. To explore this further, we ran a small simulation study to determine the effect the shape of the true underlying distribution had on the direction of the bias of the lognormal estimator.

I generated data from different known distributions for the non-zero component of the population in order to see if there was a relationship between the shape of the distribution and the direction of the bias of the lognormal portion of the estimator of the delta-lognormal mean (derivation from Finney (1941); see also Bradu & Mundlak, 1970). Several distributions were used to vary skew and kurtosis (Table 6). For each distribution, we simulated 500,000 observations and calculated the lognormal estimator of the mean (Table 6). The skewnormal, gamma, and exponential distributions have moderate skew compared to the lognormal distribution with similar mean values; they show some evidence of bias but the bias is generally small and positive. On the other hand, the estimator based on a log-exponential distribution was severely negatively biased for distributions that had very long tails to the right with high

Table 5. Results of 25000 simulations of random samples from one of two distributions contaminated with additional observations of 0 for a total sample size of n = 125. Estimation is done based either on the delta-lognormal approach or the usual sample mean and standard error of the mean estimators. The distributions for $log(X)$ are normal or exponential with a mean and variance of 1.

(a) 80% non-zeroes (m = 100 non-0s; r = 25 0s)

Distribution	Method	Estimated Mean	Estimated SEM	Coverage of 95% CI[@]
Lognormal	Δ-lognormal	3.5860	0.4601	0.9486
Exponential	Δ-lognormal	3.6475	0.4816	0.7557
Lognormal	$\bar{x}$	3.5872	0.4751	0.9327
Exponential	$\bar{x}$	11.6888	7.8066	0.9574

(b) 40% non-zeroes (m = 50 non-0s; r = 75 0s)

Lognormal	Δ-lognormal	1.7916	0.3585	0.9570
Exponential	Δ-lognormal	1.8581	0.3894	0.8065
Lognormal	$\bar{x}$	1.7925	0.3639	0.9444
Exponential	$\bar{x}$	5.8432	3.7457	0.9348

(c) 20% non-zeroes (m = 25 non-0s; r = 100 0s)

Lognormal	Δ-lognormal	0.8965	0.2603	0.9454
Exponential	Δ-lognormal	0.9726	0.3067	0.8040
Lognormal	$\bar{x}$	0.8968	0.2608	0.9322
Exponential	$\bar{x}$	2.9217	1.9198	0.8832

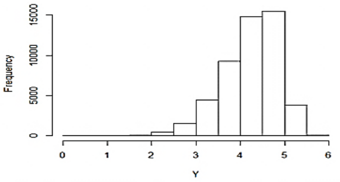
[@] Caution must be taken in comparing these coverage rates to those in Table 3. Here the number of zero observations is fixed and so has no sampling variability to contribute to the overall variance of the estimators. As a result, the coverage rates more appropriately show the behavior of the non-zero portion of the estimators.

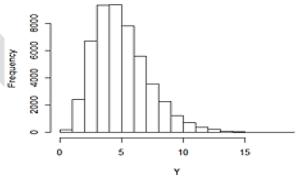
skew. This result is not due to large means or variances since the lognormal estimator of the mean is always unbiased for any lognormal distribution, even those with large means and variances. Hence, these results imply that Finney's estimator, and hence the delta-lognormal estimator, is a reasonable choice as long as the right tail of the observed data is not too long relative to that expected for a lognormal distribution with a similar mean. Unfortunately, because the direction of the bias depends on the shape of the true underlying distribution, use of the delta-lognormal approach is difficult in many situations without strong belief that the data truly are selections from a lognormal distribution.

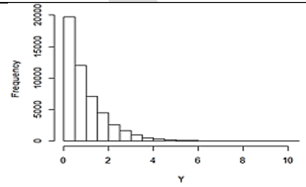
Comment: while not the subject of this manuscript another result is evident from these simulations. The alternative to using the delta-lognormal estimator when the non-zero data are not lognormally distributed and the actual distribution of the non-zero data is unknown is to use the sample mean and sample standard deviation. One might be tempted then to use large-sample theory for developing a confidence interval, i.e. assume the sample size is sufficiently large that approximate normality of the sample mean is obtained. Unfortunately, even with the sample sizes we had here (25, 50 and 100 non-zero values), the sampling distribution of the sample mean is skewed right when the non-zero data are highly skewed. Hence, as the number of non-zero observations in the sample decreases, the coverage of the large-sample CI also decreases. We recommend bootstrapping using the same sampling design and the appropriate assumptions about whether the population is infinite or finite to obtain bootstrap-t confidence intervals that have better coverage than the large sample theory CI.

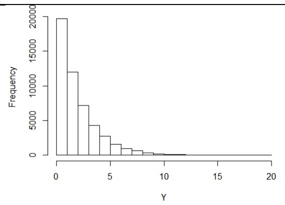
Table 6. Expected value of the Finney's (1947) lognormal estimator of the mean for three distributions.

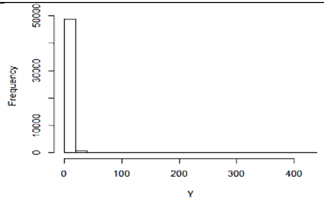
Distribution	$E[Y]$	Expected Value of Finney's LN estimator	Relative Bias $\frac{E_{LN}[Y]}{E[Y]}$	Shape of Distribution

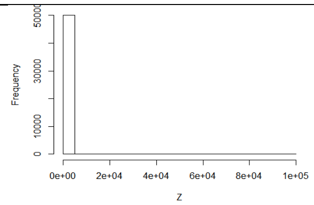
Y ~ Skewnormal(5,1,-4)
Var = 0.40
Skew = –0.79
Kurtosis = 0.66
	4.227	4.231	1.009	

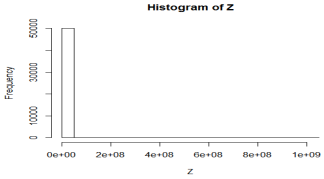
Y ~ Gamma(5,1)
Var = 5.00
Skew = 0.89
Kurtosis = 1.19
	5.000	5.039	1.008	

Y ~ Exponential(1)
Var = 1.00
Skew = 2.00
Kurtosis = 6.04
	1.000	1.282	1.282	

Y ~ Exponential(0.5)
Var = 4.00
Skew = 2.00
Kurtosis = 5.87
	2.000	2.553	1.277	

$Ln(Y)$ ~ Lognormal(1,1)
Var = 32.28
Skew = 6.10
Kurtosis = 76.53
	4.482	4.482	0	

$Ln(Y)$ ~ Exponential(1)
Var = 501782.3
Skew = 549.05
Kurtosis = 343926.9
	11.686	4.462	0.382	

$Ln(Y)$ ~ Exponential(0.75)
Var = 1.928055e+12
Skew = 676.67
Kurtosis = 469462.0
	2693.309	9.233	0.003	

References

Agresti, A. 2002. *Categorical Data Analysis, 2nd ed*. New York: Wiley & Sons, Inc.

Agresti, A. & Coull, B. A. 1998. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician, 52, 119-126.

Aitchison, J. 1955. On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at theOrigin. Journal of the American Statistical Association, 50, 901-908.

Aitchison, J. and Brown, J. A. C. 1957. *The Lognormal Distribution*. Cambridge: Cambridge University Press.

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association* 89, 1282-1289.

Bradu, D. & Mundlak, Y. 1970. Estimation in Lognormal Linear Models, Journal of the American Statistical Association, 65, 198-211.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Belmont.

Casella, G. & Berger, R. L. 2002. *Statistical Inference, 2nd ed*. Pacific Grove, CA: Duxbury

Cass-Calay, S. L. 2010. Standardized catch rates of large bluefin tuna (*Thunnus thynnus*) from the U.S. pelagic longline fishery in the Gulf of Mexico during 1987-2009. Collect. Vol. Sci. Pap. ICCAT, 66(3): 1257-1267.

Cochran, W. G. 1977. *Sampling Techniques, 3rd ed*. New York: John Wiley & Sons, Inc.

Cook, J. D. 2008. Illustrating the error in the delta method. Available at http://www.johndcook.com/delta_method.pdf.

Davison, A. C. & Hinkley, D. V. 1997. Bootstrap Methods and their Application. Cambridge University Press.

Efron, B. and Tibshirani, R. (1986). The Bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, Vol 1., No. 1, pp 1-35.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.

Finney, D. J. 1941. On the distribution of a variate whose logarithm is normally distributed, Journal of the Royal Statistical Society, Series B, 7, 155-61

Glasser, G. J. 1961. An Unbiased Estimator for Powers of the Arithmetic Mean, *Journal of the Royal Statistical Society. Series B (Methodological),* 23, 154-159.

Goldberger, A. S. 1968. The Interpretation and Estimation of Cobb-Douglas Functions. Econometrica, 36, 464-472.

Goodman, L. A. 1960. On the exact variance of products. Journal of the American Statistical Association, 55(292), 708 – 713.

Gregoire, T. G. & Schabenberger, O. 1999. Sampling-Skewed Biological Populations: Behavior of Confidence Intervals for the PopulationTotal. Ecology, 80, 1056-1065.

Hastie and Tibshirani (1990) *Generalized Additive Models*. Chapman and Hall.

Hinkley, D.V. and Shi, S. (1989), Importance sampling and the nested bootstrap, *Biometrika,* 76, 435-446.

Hoyle, M. H. 1968. The estimation of variances after using a Gaussianating transformation. The Annals of Mathematical Statistics, 39, 1125-1143.

Ingram, G. W., Richards, W. J., Porch, C. E., Restrepo, V., Lamkin, J. T., Muhling, B., Lyczkowski-Shultz, J., Scott, G. P., and Turner, S. C. 2008. Annual Indices of Bluefin Tuna (*Thunnus Thynnus*) Spawning Biomass in the Gulf of Mexico Developed Using Delta-Lognormal and Multivariate Models. *SCRS/2008/086*, 30 pg.

Ingram Jr., J. W., Richards, W. J., Lamkin, J. T., and Muhling, B. 2010. Annual indices of Atlantic bluefin tuna (Thunnus thynnus) larvae in the Gulf of Mexico developed using delta-lognormal and multivariate models. Aquat. Living Resour. 23, 35–47

Ingram, G. W. 2011. Delta-lognormal and multinomial approaches to index development for parrotfish, silk snapper, and queen snapper from Puerto Rican Trip Tickets. SEDAR26 – DW – 07. 17 pg.

IATTC (Inter-American Tropical Tuna Commission), 2007. Standardization of yellowfin and big eye tuna CPUE data from Japanese longliners, 1975 – 2004. Working Group To Review Stock Assessments, 7th Meeting, La Jolla, California (USA), 15-19 May 2006, Document SAR-7-07. 19 pg.

Lehmann, E. L. 1983. *Theory of Point Estimation*. New York: John Wiley & Sons, Inc.

Lo, N. C. H., L. D. Jacobson, and J. L. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Can. J. Fish. Aquat. Sci. 49: 251 5-2526.

MacCall, A. D. 1990. Dynamic geography of marine fish populations. University of Washington Press, Seattle. 153 pp.

McCarthy, P. J. & Snowden, C. B. 1985 The bootstrap and finite population sampling. Vital and Health Statistics, Public Health Service, US Dept of Health and Human Services. NCHS Series 2, No. 95, 23 pp.

McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models. 2nd ed*. London: Chapman & Hall.

Mohammadi, M., Salehi, M., and J. N. K. Rao, 2014. Bootstrap confidence intervals for adaptive cluster sampling design based on Horvitz–Thompson type estimators. Environ. Ecol. Stat., 21, 351–371.

Myers, R. A. & Pepin, P. 1990. The Robustness of Lognormal-Based Estimators of Abundance. Biometrics, 46, 1185-1192.

Neter, J., Wasserman, W., and Kutner, M. H. 1990. *Applied Linear Statistical Models, 3rd ed.* Homewood, NJ: Richard D. Irwin, Inc.

Newton, M. A. & Geyer, C. J. 1994. Bootstrap Recycling: A Monte Carlo Alternative to the Nested Bootstrap. Journal of the American Statistical Association, 89, 905-912.

Owen, W. J. & DeRouen, T. A. 1980. Estimation of the Mean for Lognormal Data Containing Zeroes and Left- Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants. Biometrics, 36, 707-719.

Pennington, M. 1983. Efficient Estimators of Abundance, for Fish and Plankton Surveys. Biometrics, 39, 281-286.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association 83, 231-241.

Rodriguez – Martin et al 2003.

Shono, H. 2008. Confidence interval estimation of CPUE year trend in delta-type two-step model, Fisheries Science, 74, 712–717.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association* 87, 775-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics 20,* 135-154.

Smith, S. J. 1988. Evaluating the Efficiency of the Δ-Distribution Mean Estimator. Biometrics, 44, 485-493.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science, 53, 577–588.

Syrjala, S. E. 2000. Critique on the use of the delta distribution for the analysis of trawl survey data. ICES Journal of Marine Science, 57, 831–842.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso, JRSSB, 58, 267–288.

Thompson, S. K. 2002. Sampling, 2nd ed. New York: John Wiley & Sons, Inc.

Walter, J. & Ortiz, M. 2012. Derivation of the delta-lognormal variance estimator and recommendation for approximating variances for two-stage cpue standardization models. Collect. Vol. Sci. Pap. ICCAT, 68, 365-369.

Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673-686.

Wood, S.N. 2006. Generalized Additive Models: an introduction with R, CRC

Zou, G. Y., Taleban, J., and Huo, C. Y. 2009. Confidence interval estimation for lognormal data with application to health economics. Computational Statistics and Data Analysis, 53, 3755-3764.