

Abstract.—The spatial distribution of marine organisms is highly patchy. Because of this patchy distribution, data from marine abundance surveys are highly skewed and have a large variance. Compounding the problem of estimating the mean abundance from such data, is that occasionally a relatively huge catch will occur. These large catches are not “outliers” but do dominate the estimates of the mean and variance. A lognormal model of the non-zero survey values (a Δ -distribution) is used to model survey data. The estimators, based on the lognormal model, appear to be much more efficient for marine data than the usual sample estimators. In particular, the lognormal-based estimators provide reasonable estimates for data sets that contain a very large catch. The properties and efficiency of the Δ -distribution estimators are examined and the techniques are applied to various marine data sets.

Estimating the mean and variance from highly skewed marine data

Michael Pennington

Woods Hole Laboratory
National Marine Fisheries Service, NOAA
Woods Hole, Massachusetts 02543

Characteristically, the observed distribution of abundance data generated by marine surveys has a large variance, is highly skewed to the right, and contains a substantial proportion of zeros. Because of this large variability, the sample mean has a low level of precision even for relatively intensive surveys (Grosslein, 1971; Godø, 1994; Pennington and Godø, 1995). A common problem in the analyses and interpretation of skewed survey data, is that a single immense catch may account for 50% or more of the total catch during a survey (Sissenwine, 1978; Dew, 1990; McConnaughey and Conquest, 1992; Bowering and Brodie, 1994). These extreme values not only greatly affect the estimate of the mean but also of the variance (Otto, 1986). As McConnaughey and Conquest (1992) observed, although these large values cause much uncertainty for management, they reflect the spatial distribution of the species and are not outliers that should be discarded. In practice, the use of more efficient sampling schemes or estimators is the only realistic way to increase survey precision; the total number of samples that can be taken is limited by the high cost of sampling at sea (Gunderson, 1993).

One possible way to increase the precision of survey estimates is to model the observed distribution of catches and exploit the model's properties to develop more efficient estimators of population param-

eters (see, e.g. Pennington, 1983; MacLennan and MacKenzie, 1988; Lo et al., 1992; McConnaughey and Conquest, 1992; Conquest et al., 1996; Stefánsson, in press). For marine data, the distribution of the nonzero values is often well approximated by a lognormal distribution (e.g. Pennington, 1983; Smith, 1988; McConnaughey and Conquest, 1992; Conquest et al., 1996). Myers and Pepin (1990) found that of the 69 marine data sets they examined, only 5 differed significantly from the lognormal distribution. Thus the lognormal model has been used as a basis for developing survey abundance estimators (e.g. Pennington, 1983, 1986; Lo et al., 1992; McConnaughey and Conquest, 1992) and for estimating commercial catch (Conquest et al., 1996).

It is not surprising that marine abundance data often appear to follow a lognormal distribution. The factors that determine abundance over a region seem to have a multiplicative effect. When this is the case, survey data will be approximately lognormally distributed by the central limit theorem (see, e.g. Aitchison and Brown, 1957). More generally, the lognormal model has been useful for analyzing a wide range of ecological data. As Dennis and Patil (1988) put it: “Ecological abundance data are intrinsically positive, with a few enormously high data points typically arising in every study. The lognormal distribution is an ideal descriptor of such

data with a positive range, right skewness, heavy tail, and easily computed parameter estimates."

To estimate efficiently the mean of skewed marine survey data and to be able to assess its precision, I examined an estimator based on a lognormal model of the distribution. I present the estimator's theoretical efficiency, assess its performance by applying it to several real marine data sets, and give methods for constructing confidence intervals.

Statistical methods

Suppose the nonzero catches generated by a survey are lognormally distributed, i.e. the logged values are normally distributed. If the distribution contains a proportion of zeros, then it is called a Δ -distribution (Aitchison and Brown, 1957). If zeros do not occur, then it is the usual lognormal distribution.

Estimating the mean and variance of the Δ -distribution

As is the case for any distribution, the sample average, \bar{x} , and variance, s_x^2 , are unbiased estimators of the mean and variance of the Δ -distribution. Because of the properties of the lognormal distribution, the minimum variance unbiased estimators (denoted by c and d) of the mean and variance of the Δ -distribution are given by (Aitchison and Brown, 1957)

$$c = \begin{cases} \frac{m}{n} \exp(\bar{y}) g_m(s^2/2), & m > 1 \\ \frac{x_1}{n}, & m = 1 \\ 0, & m = 0 \end{cases} \quad (1)$$

and

$$d = \begin{cases} \frac{m}{n} \exp(2\bar{y}) \left\{ g_m(2s^2) - \left(\frac{m-1}{n-1} \right) g_m\left(\frac{m-2}{m-1} s^2\right) \right\}, & m > 1 \\ \frac{x_1^2}{n}, & m = 1 \\ 0, & m = 0 \end{cases} \quad (2)$$

where n is the number of observations, m is the number of nonzero values, $y = \ln(x)$, \bar{y} , and s^2 are the sample mean and variance of the logged nonzero values, x_1 denotes the single untransformed value when m equals one, and $g_m(t)$, which is a function of m and t (e.g. $t = s^2/2$ in Equation 1), is defined by

$$g_m(t) = 1 + \frac{m-1}{m} t + \sum_{j=2}^{\infty} \frac{(m-1)^{2j-1}}{m^j(m+1)(m+3)\dots(m+2j-3)} \times \frac{t^j}{j!} \quad (3)$$

Estimating the variance of \bar{x} and c

Again for the Δ -distribution, the sample mean, \bar{x} , and c are both unbiased estimators of the mean. Likewise, the sample variance, s_x^2 , and d are unbiased estimators of the population variance. If \bar{x} is used to estimate the mean, then s_x^2/n , the sample variance divided by the sample size, is an estimate of the variance of \bar{x} . But s_x^2 can be a very inefficient estimator compared with d , and, therefore, it is frequently recommended that d/n be used to estimate the variance of \bar{x} (Aitchison and Brown, 1957). The minimum variance unbiased estimator of the variance of c is given by (Pennington, 1983)

$$\text{var}_{est}(c) =$$

$$\begin{cases} \frac{m}{n} \exp(2\bar{y}) \left\{ \frac{m}{n} g_m^2(s^2/2) - \left(\frac{m-1}{n-1} \right) g_m\left(\frac{m-2}{m-1} s^2\right) \right\}, & m > 1 \\ \left(\frac{x_1}{n} \right)^2, & m = 1 \\ 0, & m = 0 \end{cases} \quad (4)$$

If $m = n$, i.e. there are no zeros, then Equations 1, 2, and 4 reduce to the usual estimators for the lognormal distribution.

Relative efficiency of \bar{x} and c

For the two estimators of the mean, \bar{x} and c , the one with the smallest variance is the most efficient estimator. The formulas in the last section give estimates of the variance based on the particular sample drawn from the distribution. The expected or true variance of \bar{x} is (Aitchison and Brown, 1957)

$$\text{var}(\bar{x}) = \frac{\exp(2\mu + \sigma^2)}{n} \{ p[\exp(\sigma^2) - 1] + p(1-p) \}, \quad (5)$$

where μ is the mean and σ is the standard deviation of the log-transformed nonzero values, and p is the proportion of nonzeros. Smith (1988) derived the expected value of $\text{var}_{est}(c)$, which, in the same notation as above, is given by

$$\text{var}(c) = \frac{\exp(2\mu + \sigma^2)}{n} \times \left\{ \frac{1}{n} E_{m>0} \left[m^2 \left\{ \exp(\sigma^2/m) g_m(\sigma^4/2m) - 1 \right\} \right] + p(1-p) \right\}, \quad (6)$$

where

$$E_{m>0}[\cdot] = \sum_{m=1}^n m^2 \binom{n}{m} p^m (1-p)^{n-m} \times \left\{ \exp(\sigma^2/m) g_m(\sigma^4/2m) - 1 \right\}.$$

It can be shown using results from Bradu and Mundlak (1970) that the $\text{var}(c)$ is always less than or equal to $\text{var}(\bar{x})$, both decrease as n increases, but $\text{var}(c)$ decreases more quickly than does $\text{var}(\bar{x})$. For values of σ^2 typical for marine data, $\text{var}(c)$ is considerably smaller than $\text{var}(\bar{x})$ (Pennington, 1986; Smith, 1988). This can be seen in Figure 1 which contains plots of $\text{var}(c)$ divided by $\text{var}(\bar{x})$ versus sample size for a range of σ^2 's appropriate for marine survey data.

Tracking trends in abundance

For a series of marine surveys, it is usually assumed that the mean catch per tow is proportional to population size. If this is the case, then the estimator, c ,

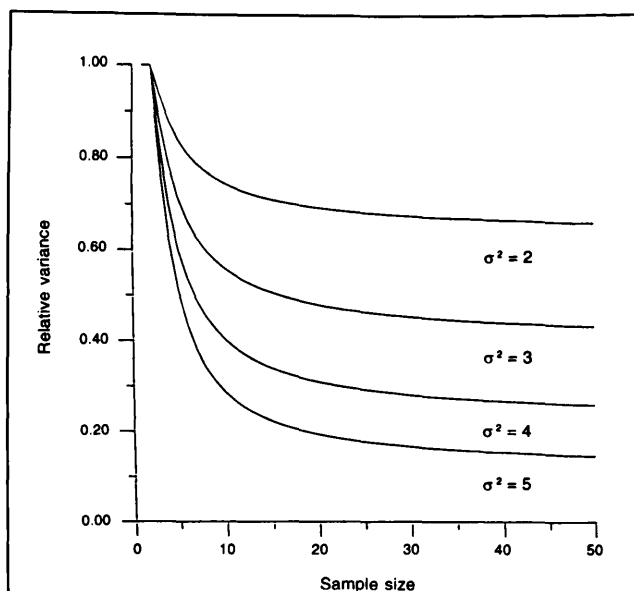


Figure 1

The relative efficiency of the estimators c and \bar{x} for estimating the mean of the Δ -distribution. The plots show the $\text{var}(c)$ divided by the $\text{var}(\bar{x})$ as a function of sample size when the variance of the nonzero logged values, σ^2 , equals 2, 3, 4, and 5, and when the proportion of nonzero values, p , is 0.8.

is an index of abundance. The mean of the lognormal distribution is given by $\exp(\mu + \sigma^2/2)$. McConaughy and Conquest (1992) have suggested that for lognormally distributed survey data, $\exp(\bar{y})$, a slightly biased estimate of $\exp(\mu)$, may be a more stable index for following trends in abundance than estimates of the mean. That is, if σ^2 is constant over time (which is equivalent to the coefficient of variation of the untransformed variable being constant) then $\exp(\mu)$, the median of the lognormal distribution, will also be proportional to abundance. The variance of $\exp(\bar{y})$ can be considerably smaller than the variance of c .

The mean of the Δ -distribution is $p[\exp(\mu + \sigma^2/2)]$. If the mean is proportional to population size and σ^2 is constant for a survey series, then $p[\exp(\mu)]$ will also be an index of abundance. It can be shown with techniques similar to those in Pennington (1983) that the minimum variance unbiased estimator, k , of $p[\exp(\mu)]$ is

$$k = \begin{cases} \frac{m}{n} \exp(\bar{y}) g_m\left(\frac{-s^2}{2(m-1)}\right), & m > 1 \\ \frac{x_1}{n}, & m = 1 \\ 0, & m = 0 \end{cases} \quad (7)$$

and the minimum variance unbiased estimator of the variance of k is given by

$$\text{var}_{est}(k) = \begin{cases} \frac{m}{n} \exp(2\bar{y}) \left\{ \frac{m}{n} g_m^2\left(\frac{-s^2}{2(m-1)}\right) - \left(\frac{m-1}{n-1}\right) g_m\left(\frac{-2s^2}{m-1}\right) \right\}, & m > 1 \\ \left(\frac{x_1}{n}\right)^2, & m = 1 \\ 0, & m = 0 \end{cases} \quad (8)$$

As before, if $m = n$, then Equations 7 and 8 reduce to the lognormal case (Bradru and Mundlak, 1970).

Confidence intervals

If n is large, then $c \pm 2[\text{var}_{est}(c)]^{1/2}$ and $k \pm 2[\text{var}_{est}(k)]^{1/2}$ are approximately 95% confidence intervals. For smaller n , a conservative approach for constructing confidence intervals is to calculate separate in-

tervals for p and for the mean (or median) of the lognormally distributed nonzero values. For example, if (p_L, p_U) is a 95% confidence interval for p and (L, U) is a 95% interval for the mean (or median) of the lognormal component (see, e.g. McConnaughey and Conquest, 1992), then $(p_L L, p_U U)$ will have a confidence level of at least 90% ($=0.95 \times 0.95$).

Examples

There are two types of data sets that are typical for marine abundance surveys. The first type has a single large catch that can be many times larger than the next biggest catch. This huge catch may account for more than 50% of the total catch taken during the survey. The other category, and the more common type, is that the distribution of catches is highly skewed, as is the case for the first type, but there are no isolated large catches that dominate the total catch. These are the basic types of data sets that would be expected if samples are taken from a highly skewed lognormal distribution.

Isolated large catches

Occasionally, a very large value can occur when samples are drawn from a lognormal distribution. The first example (Table 1) is an artificial data set

generated from a lognormal distribution with $\mu = 0$ and $\sigma^2 = 4$. The mean of the distribution is 7.4 and its variance is 2,926. Because of one large point in the sample, the estimates, $\bar{x} = 38.8$ and $s_x^2 = 63,320$, are much larger than the true values. The estimated standard error of the sample mean based on the sample variance is 35.6 $[= (63,320/50)^{1/2}]$.

The sample estimates of the logged values are $\bar{y} = 0.175$ and $s^2 = 3.921$. Hence the estimates of the mean and variance from the minimum variance unbiased estimators are [Equations 1 and 2, $m = n = 50$]

$$c = \exp(0.175)g_{50}(1.961) = 7.6$$

and

$$d = \exp(0.350) \left\{ g_{50}(7.842) - g_{50}\left(\frac{48}{49} \times 3.921\right) \right\} \\ = 1.42 \times (922.83 - 34.07) = 1261,$$

which are much closer to the true values than are the ordinary sample estimates. The estimate of the standard error of the sample mean using d is 5.0 $[=(1261/50)^{1/2}]$ as compared with an estimate of 35.6 based on the sample variance. The expected standard error of the sample mean (when $n = 50$) is 7.6 $[=(2926/50)^{1/2}]$.

The estimated variance of c is given by (Equation 4)

Table 1

An artificial data set generated from a lognormal distribution, catch per unit of effort (CPUE) data for red king crab from a 1991 trawl survey in the Bering Sea, and CPUE data for petrale sole from a 1992 survey off the west coast of the United States.

Artificial data

0.03	0.03	0.04	0.09	0.14	0.18	0.20	0.22	0.25	0.25
0.25	0.27	0.27	0.27	0.30	0.38	0.40	0.42	0.51	0.55
0.61	0.74	0.80	0.82	0.84	0.86	1.02	1.41	1.91	2.08
2.35	2.72	3.20	3.21	3.22	3.26	3.39	4.21	4.76	5.61
7.47	8.06	8.25	8.80	8.85	9.10	12.92	16.94	23.61	1782.19

Red king crab CPUE¹

65	66	69	71	72	73	76	78	78	79
81	83	84	85	86	108	141	144	152	154
154	157	160	161	165	183	204	234	263	265
292	313	317	339	413	435	455	469	569	609
625	746	758	813	909	1056	1145	1197	1339	1789
1842	2520	4194	4286	5424	32538				

Petrale sole CPUE²

4.89	6.31	7.66	11.12	11.31	12.23	13.16	13.61	20.96	23.64
26.31	26.79	28.58	30.07	43.87	56.49	71.43	81.08	94.30	103.02
112.71	115.48	153.96	169.52	179.83	199.80	205.71	310.82	398.51	4635.18

¹ The data set also includes 24 zeros.

² Ten zeros are not shown.

$$\begin{aligned}\text{var}_{\text{est}}(c) &= \exp(0.350)\{g_{50}^2(1.961) - g_{50}(3.841)\} \\ &= 1.42\{(6.39)^2 - 34.07\} = 9.7.\end{aligned}$$

Thus the estimated standard error of c is 3.1 ($= (9.7)^{1/2}$) as compared with an expected value of 3.8 (from Equation 6, $p = 1$). For further examples of the performance of the estimators on lognormal data, see Aitchison and Brown (1957), Blackwood (1991), McConnaughey and Conquest (1992) and Conquest et al. (1996).

The next three examples are survey data that are similar in appearance to the artificial data set in that each contains a single large isolated catch. The first data set is from a trawl survey in the southeastern Bering Sea in 1991. In Table 1 is shown the survey catch per unit of effort (CPUE) of male red king crab, *Paralithodes camtschatica*, of legal size. The largest CPUE is six times greater than the second largest value and accounts for nearly 50% of the total survey catch. In Table 2 the estimates of the mean and the standard errors are calculated as above. The pattern of the estimates is similar to that for the artificial data. In particular, the estimate of the mean, $c = 545.0$, is much smaller than the sample mean (864.8) and appears to be more precise. The reason that c is so much lower than \bar{x} is that, based on the Δ -lognormal model of the data, a CPUE as large or larger than the biggest value (32,538) would have occurred for approximately 1 in 2,200 tows during the 1991 crab survey, and c weighs the value accordingly. In contrast, the sample mean gives each CPUE value equal weight.

In Table 1, CPUE data are given for petrale sole, *Eopsetta jordani*, from a 1992 trawl survey off the west coast of the United States. The largest catch is 65% of the total catch and is 12 times larger than the next largest catch. Estimates of the mean and standard errors are given in Table 2.

The last example of this type of data set is from a trawl survey off the east coast of the United States. The data (Sissenwine, 1978) are the catch per tow in 1973 of Atlantic mackerel, *Scomber scombrus*. The largest catch (5,182 kg) is more than 25 times greater than the next largest (194 kg) and is 92% of the total catch. This is the one example presented for which lognormality of the nonzero values was rejected ($P=0.02$). Though the estimate $c = 2.0$ kg/tow is considerably smaller than the sample mean ($\bar{x}=26.2$ kg/tow), it is much more consistent with previous and subsequent survey indices (e.g. 1.6 kg/tow in 1972 and 2.5 kg/tow in 1974) than is the sample mean (see Fig. 5 in Sissenwine, 1978).

No dominating large catch

The more usual type of survey data set is one that is highly skewed but does not contain a relatively large isolated value. A typical example of this sort of data is seen in Figure 2 which shows the catch per tow of juvenile Arcto-Norwegian cod, *Gadus morhua*, collected during a 1989 midwater trawl survey in the western Barents Sea (Helle, 1994). The estimate of the mean from c is 55.2 and from \bar{x} is 49.7. Similarly, the estimate d is greater than the sample variance (Table 2).

Another example is from a 1989 zooplankton survey in the Barents Sea (Helle, 1994). Figure 3 is a plot of the biomass per tow of copepods sampled with a Juday plankton net. The frequency distribution is similar to that in Figure 2, and, again, the estimates c and d are larger than the ordinary sample estimates (Table 2).

The sample average and variance will be underestimates for most samples (i.e. be smaller than the true values). This is due to the sampling distribution of \bar{x} and s_x^2 , which, for a highly skewed distribution, will still be skewed to the right for small to

Table 2

Summary statistics for estimating the mean and standard errors for six data sets, where n is the sample size, m is the number of nonzeros, \bar{x} is the sample mean, se_s is the sample standard error (SE), se_d is the SE based on a lognormal model, \bar{y} and s^2 are the mean and variance of the nonzero logged values, c is the estimate of the mean based on a lognormal model, and $[\text{var}_{\text{est}}(c)]^{1/2}$ is its estimated standard error. The expected or true values for the artificial data are in parentheses.

Data set	n	m	\bar{x}	se_s	se_d	\bar{y}	s^2	c	$[\text{var}_{\text{est}}(c)]^{1/2}$
Artificial	50	50	38.8 (7.4)	35.6 (7.6)	5.0 (7.6)	0.175 (0)	3.921 (4)	7.6 (7.4)	3.1 (3.8)
Red king crab	80	56	864.8	415.0	156.0	5.755	1.866	545.0	134.8
Petrable sole	40	30	179.2	115.2	46.9	3.96	2.238	112.1	39.8
Atlantic mackerel	216	67	26.2	24.0	1.4	-0.165	4.269	2.0	0.8
Juvenile cod	161	99	49.7	11.7	26.7	2.759	3.572	55.2	16.4
Zooplankton	160	160	478.1	68.9	93.2	5.338	1.876	525.7	77.2

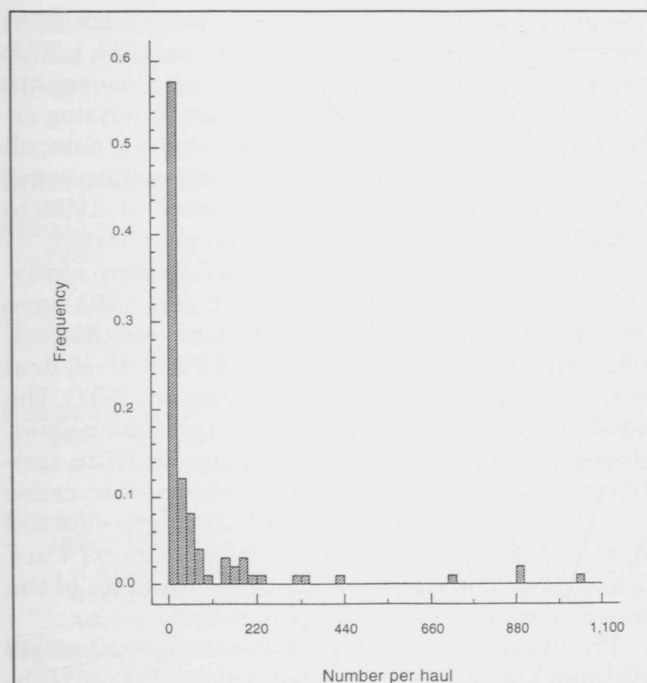


Figure 2

Frequency plot of the number per tow of juvenile Arcto-Norwegian cod from a 1989 midwater trawl survey in the Barents Sea (Helle, 1994). There were also 62 zeros ($n=161$) which are not included in the plot.

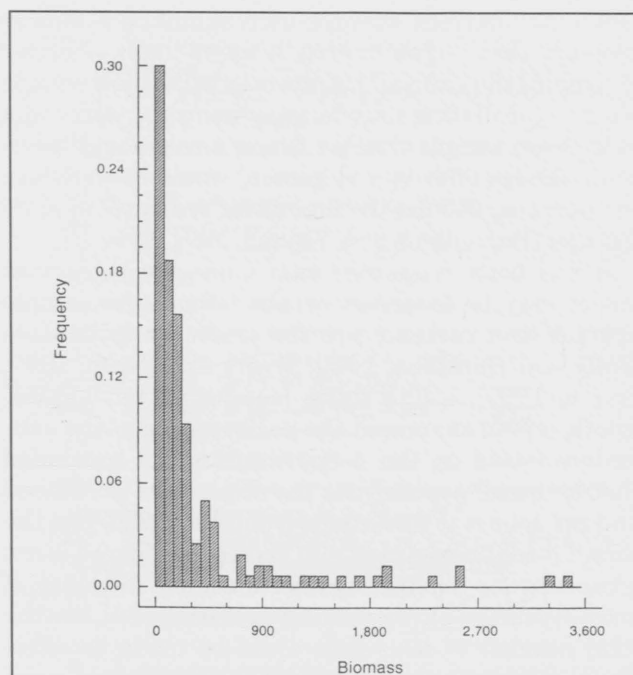


Figure 3

Biomass ($\mu\text{g}/\text{m}^3$) per tow of copepods *Calanus finmarchicus* and *Oithona similis*. Data are from a 1989 plankton survey in the Barents Sea (Helle, 1994). There were no zero values ($n=160$).

moderate sample sizes. The median can be much smaller than the mean for a skewed distribution, and, therefore, the sample estimators are not only less efficient but will underestimate the true values of the parameters most of the time (Pennington, 1983; McConnaughey and Conquest, 1992; Conquest et al., 1996). The sampling distribution of s_x^2 is much more skewed than is the sampling distribution of \bar{x} , which is the reason that s_x^2 often greatly underestimates its expected value more often than does \bar{x} (Pennington, 1986). The sample estimators are unbiased, even though most of the time the estimates are low and are very high for the occasional sample that contains a huge catch (McConnaughey and Conquest, 1992).

Discussion

The estimators of abundance, based on the lognormal model, perform as expected on real survey data if the underlying model for the nonzero values is a lognormal distribution. The estimates are more precise, and the occasional huge catch does not affect the estimates nearly as much as it does the sample average (see also McConnaughey and Conquest,

1992). The Δ -estimators treat these large catches as part of the distribution, as a reflection of how fish are actually distributed spatially, eliminating the need to handle them as "outliers," that is, to discard the points arbitrarily in an analysis of the data. Since all models only approximate reality, an advantage in using lognormal-based estimators for marine data is that they appear to be fairly robust to deviations from the model (Blackwood, 1991; Pennington, 1991; Conquest et al., 1996).

The Δ -estimators can be much more efficient than the sample estimators but lose this advantage for small samples (see Smith, 1988; Fig. 1). Thus for stratified surveys in which the region is divided into many relatively small strata and only a few stations are selected in each stratum, little would be gained by using the Δ -estimators (Smith, 1988). Only a slight gain in precision is usually achieved by increasing the number of strata beyond 6 (Cochran, 1977). Consequently it appears that a better survey design would be one that has larger strata with at least 20–30 stations in each stratum (Fig. 1). Not only would this design improve the efficiency of the Δ -estimators but it would then be possible to exploit optimal sample allocation schemes that may be more efficient (Gavaris and Smith, 1987; Polacheck and Vølstad,

1993). For current surveys with sampling intensity proportional to stratum area, it would likely be better to combine the small strata into ones with larger sample sizes for calculating abundance estimates. Another way to increase sample sizes for future surveys and to improve survey efficiency in general would be to reduce tow duration and use the time saved to sample at more stations (Pennington and Vølstad, 1991, 1994).

It has been suggested that since the lognormal model may be incorrect or not robust, the sample average and variance are the preferred estimators (Jolly and Hampton, 1990; Myers and Pepin, 1990; Smith, 1990). Using finite population techniques, Smith (1990) examined the performance of the estimators based on the Δ -distribution and concluded that for small populations the estimators are biased and not robust to deviations from the model. But the sort of model-based bias that Smith considered is not a concern for marine surveys. Because for most, if not all, marine surveys, the population size, i.e. the total number of tows that could be made, is effectively infinite, whereas Smith's simulations were samples from populations of size 30. There is no reason that the Δ -estimators should be unbiased if applied to samples from small populations. For Smith's simulations, the usual properties of the lognormal-based estimators are apparent if the small samples ($n=3$) are assumed to be from a larger population. That is, if the samples of size 3 are assumed to come from a survey for which the possible number of tows (the population size) is large, then the estimators are unbiased (see Table 1 in Smith, 1990). The model-based bias that Smith observed is a function of population size, not a property as such of the Δ -estimators or the size of the sample.

What would cause concern is the possibility that the underlying distribution may have appeared to be approximately lognormal but was not and that the departure from lognormality caused the lognormal-based estimates and inferences to be misleading. Myers and Pepin (1990) have claimed, motivated by some simulations, that lognormal-based estimators are very sensitive to undetectable deviations from lognormality. But to test a model fairly, the alternative models should be realistic. The nonrobustness that they observed was simply due to the contamination of lognormal distributions with very small values, the opposite of what causes the imprecision of abundance estimates from marine surveys, i.e. the large catches (Pennington, 1991). It was not only that the contaminating values were small, but there was a relatively high probability that small values would occur. Since $\ln x$ goes to minus infinity as x approaches zero, these small values resulted in large negative values on the log scale, which caused the ex-

treme instability of the lognormal-based estimators in Myers and Pepin's simulations. Aitchison (1986, p. 270) made the same point when discussing a sensitivity analysis of another log-based procedure. Analyzing artificial data is no different from analyzing real data; all aspects of the simulated data should be examined carefully (see, e.g. McConnaughey and Conquest, 1992) to ensure that the resulting conclusions are relevant.

In practice, even if such small values were statistically "undetectable" (Myers and Pepin, 1991), one would know (e.g. by looking at the data) whether values could be arbitrarily close to zero and, if so, deal with them appropriately as in Pennington (1991). The small values that may occur after transforming abundance data for a particular length class with an age-length key (Myers and Pepin, 1991) will not cause any problems if the original catch at length data are distributed lognormally. This is because $\ln(ax) = \ln a + \ln x$, and, therefore, the log-based estimate of the mean of ax is a multiplied by that for x .

The reason most often given for using the sample estimates and not employing any modeling techniques is that the sample average and variance are always unbiased estimators (Myers and Pepin, 1990; Smith, 1990). Lognormal-based estimators may be slightly biased for some applications but they are not overly influenced by the occasional huge catches and therefore can have a considerably smaller mean square error than the sample estimates for highly skewed distributions (Conquest et al., 1996).

There are problems if the sample estimates are used for marine data (Lo et al., 1992). The estimators are very sensitive to large catches and therefore may be rather inefficient. Another difficulty is that for the sample sizes common for marine surveys, the distribution of the sample average may be far from normal for these highly skewed distributions (Sissenwine, 1978; McConnaughey and Conquest, 1992; Conquest et al., 1996). Thus the central limit theorem cannot be invoked to assess the uncertainty associated with the estimates or to make inferences. Likewise, the distribution of the Δ -estimator may not approximate a normal distribution for small samples, but for skewed distributions it appears to converge to normality more quickly than does the sample mean (Conquest et al., 1996). For small to moderate sample sizes, methods based on the lognormal model can be used to make confidence statements.

Acknowledgments

The work presented in this paper was done while I was visiting the Alaska Fisheries Science Center. The hospitality and support shown me are greatly appreciated.

I would also like to thank Russell Kappenman who gave me the data sets in Table 1 and Kristin Helle for providing me with the zooplankton and juvenile data.

Literature cited

- Aitchison, J.
1986. The statistical analysis of compositional data. Chapman and Hall, London, 416 p.
- Aitchison, J., and J. A. C. Brown.
1957. The lognormal distribution, with special reference to its uses in economics. Cambridge Univ. Press, Cambridge, 176 p.
- Blackwood, L. G.
1991. The quality of mean and variance estimates for normal and lognormal data when the underlying distribution is misspecified. *J. Chemometrics* 5:263–271.
- Bowering, W. R., and W. B. Brodie.
1994. Distribution, age and growth, and sexual maturity of American plaice (*Hippoglossoides platessoides* (Fabricius)) on Flemish Cap (NAFO Division 3M). *J. Northwest Atl. Sci.* 16:49–61.
- Bradru, D., and Y. Mundlak.
1970. Estimation in lognormal linear models. *J. Am. Stat. Assoc.* 65:198–211.
- Cochran, W. G.
1977. Sampling techniques, 3rd ed. John Wiley, New York, NY, 428 p.
- Conquest L., R. Burr, R. Donnelly, J. Chavarria, and V. Gallucci.
1996. Sampling methods for stock assessment for small-scale fisheries in developing countries. In V. F. Gallucci, S. B. Saila, D. J. Gustafson, and B. J. Rothschild (eds.), *Stock assessment: quantitative methods and applications for small scale fisheries*, p. 179–225. CRC Press, New York, NY.
- Dennis, B., and G. P. Patil.
1988. Applications in ecology. In E. L. Crow and K. Shimizu (eds.), *Lognormal distributions: theory and applications*, p. 303–330. Marcel Dekker, New York, NY.
- Dew, C. B.
1990. Behavioral ecology of podding red king crabs, *Paralithodes camtschatica*. *Can. J. Fish. Aquat. Sci.* 47:1944–1958.
- Gavaris, S., and S. J. Smith.
1987. Effect of allocation and stratification strategies on precision of survey abundance estimates for Atlantic cod (*Gadus morhua*) on the eastern Scotian Shelf. *J. Northwest Atl. Fish. Sci.* 7:137–144.
- Godø, O. R.
1994. Factors affecting the reliability of groundfish abundance estimates from bottom trawl surveys. In A. Fernø and S. Olsen (eds.), *Marine fish behaviour in capture and abundance estimation*, p. 166–199. Fishing News Books, Farnham, UK.
- Grosslein, M. D.
1971. Some observations on accuracy of abundance indices derived from research vessel surveys. *Int. Comm. Northwest Atl. Fish. Redbook* 3:249–266.
- Gunderson, D. R.
1993. Surveys of fisheries resources. John Wiley and Sons, New York, NY, 248 p.
- Helle, K.
1994. Distribution of early juvenile Arcto-Norwegian cod (*Gadus morhua* L.) in relation to food abundance and watermass properties. *ICES Mar. Sci. Symp.* 198:440–448.
- Jolly, G. M., and I. Hampton.
1990. A stratified random transect design for acoustic surveys of fish stocks. *Can. J. Fish. Aquat. Sci.* 47:1282–1291.
- 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. Aquat. Sci.* 49:2515–2526.
- McConnaughey, R. A., and L. L. Conquest.
1992. Trawl survey estimation using a comparative approach based on lognormal theory. *Fish. Bull.* 91:107–118.
- MacLennan, D. N., and I. G. MacKenzie.
1988. Precision of acoustic fish stock estimates. *Can. J. Fish. Aquat. Sci.* 45:605–616.
- Myers, R. A., and P. Pepin.
1990. The robustness of lognormal-based estimators of abundance. *Biometrics* 46:1185–1192.
1991. Reply to: On testing the robustness of lognormal-based estimators. *Biometrics* 47:1623–1624.
- Otto, R. S.
1986. Management and assessment of eastern Bering Sea king crab stocks. In G. S. Jamieson and N. Bourne (eds.), *North Pacific workshop on stock assessment and management of invertebrates*, p. 83–106. *Can. Spec. Publ. Aquat. Sci.* 92.
- Pennington, M.
1983. Efficient estimators of abundance, for fish and plankton surveys. *Biometrics* 39:281–286.
1986. Some statistical techniques for estimating abundance indices from trawl surveys. *Fish. Bull.* 84:519–525.
1991. On testing the robustness of lognormal-based abundance estimators. *Biometrics* 47:1623.
- Pennington, M., and O. R. Godø.
1995. Measuring the effect of changes in catchability on the variance of marine survey abundance indices. *Fish. Res.* 23:301–310.
- Pennington, M., and J. H. Vølstad.
1991. Optimum size of sampling unit for estimating the density of marine populations. *Biometrics* 47:717–723.
1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. *Biometrics* 50:725–732.
- Polacheck, T., and J. H. Vølstad.
1993. Analysis of spatial variability of Georges Bank haddock (*Melanogrammus aeglefinus*) from trawl survey data using a linear regression model with spatial interaction. *ICES J. Mar. Sci.* 50:1–8.
- Sissenwine, M. P.
1978. Using the USA research vessel spring bottom trawl survey as an index of Atlantic mackerel abundance. *Int. Comm. Northwest Atl. Fish. Sel. Pap.* 3:49–55.
- Smith, S. J.
1988. Evaluating the efficiency of the Δ -distribution mean estimator. *Biometrics* 44:485–493.
1990. Use of statistical models for the estimation of abundance from groundfish trawl survey data. *Can. J. Fish. Aquat. Sci.* 47:894–903.
- Stefánsson, G.
In press. Abundance indices from groundfish survey data: combining the GLM and the delta approaches. *ICES J. Mar. Sci.*