

# The Robustness of Lognormal-Based Estimators of Abundance

R. A. Myers and P. Pepin

Department of Fisheries and Oceans, Science Branch,  
P.O. Box 5667, St John's, Newfoundland A1C 5X1, Canada

## SUMMARY

We test the robustness of a method for estimating abundance that assumes that the underlying distribution of the nonzero observations is lognormal (Pennington, 1983, *Biometrics* **39**, 281-286). Violations in model assumptions that cannot reliably be detected with moderate sample sizes ( $\leq 40$ ) lead to biases and large reductions in efficiency. Unless it can be clearly demonstrated from repeated sampling that nonzero values follow a lognormal distribution, the sample mean and variance are more robust than lognormal-based estimators of mean and variance of population abundance.

## 1. Introduction

More efficient estimators of the mean and variance of the lognormal distribution exist than the sample mean and variance (Aitchison and Brown, 1957). This fact has been exploited to estimate the abundance of organisms from field surveys (Pennington, 1983). For example, the  $\Delta$ -distribution method, which assumes a lognormal distribution modified to include zero observations, has achieved extensive use in the estimation of abundance from surveys of fish and plankton populations (Pennington, 1983; Pennington and Berrien, 1984; Sherman et al., 1984; Smith, 1988). However, there is evidence that for insects (Taylor, 1984), plankton (Barnes and Marshall, 1951; Cassie, 1962; Reid, 1981) and demersal fish (Myers and Pepin, 1986; Steinarsson and Stefansson, 1986), the lognormal distribution does not always provide the best fit to population abundance data. This can be further demonstrated by an analysis of the nonzero observations from several surveys of animal abundance (Table 1), which shows that the lognormal distribution often provides no better fit to the data than either the Weibull or gamma distributions (see Figure 1 and Table 2). Results indicate that 51 of 78 samples were best fit by either Weibull or gamma distributions. Only 11, 9, and 7 samples were significantly different ( $P < .05$ ) from the Weibull, gamma, and lognormal distributions, respectively. Of the samples that were significantly different from lognormal, four were less skewed and were better fit by either Weibull or gamma distributions, and three were more skewed and better fit by the lognormal than either of the two alternative distributions. Five of these samples were significantly different from all three distributions. Some positive skewness remains after log-transformation (Table 2) because the mode is often at the smallest data category available, suggesting that the continuous lognormal distribution is not an adequate approximation to what are discrete data. In light of such findings, we must ask whether methods such as the  $\Delta$ -distribution are robust to statistically undetectable deviations from the assumption of lognormality of nonzero values.

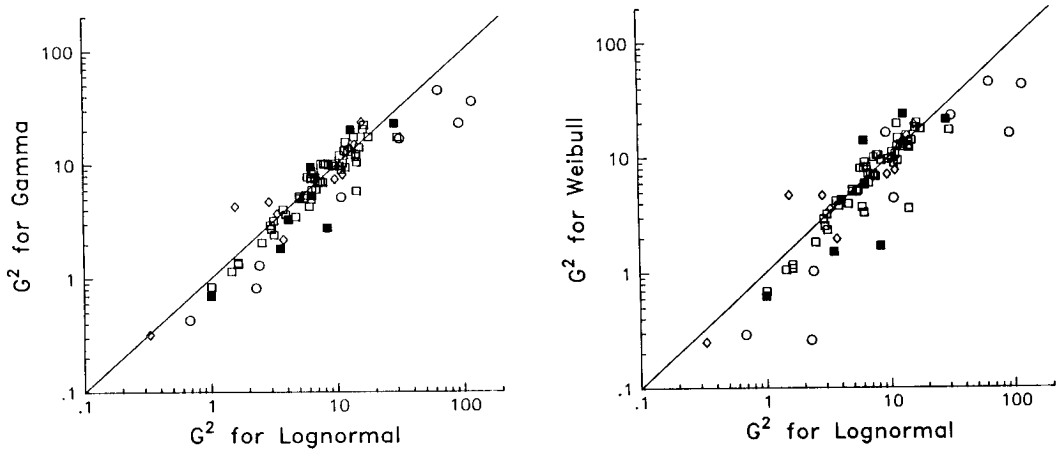
Using Monte Carlo simulations, we examine the bias and reductions in efficiency that can be expected to occur even if samples do not differ significantly from a lognormal distribution.

---

*Key words:*  $\Delta$ -distribution; Estimation of population abundance; Fish and plankton surveys.

**Table 1**  
Data sources used to fit Weibull, gamma, and lognormal distributions

Taxon	Number of samples	Sample size	Number of species	Sampling gear or unit	Source
Fish	10	23-73	2	Fishing trawl	Brodie and Wells (1985)
Polychaetes	49	14-104	18	Ekman grab	Kirkegaard (1969) Gärdefors and Orrhage (1968)
Ophiuroids	2	104	2	Ekman grab	Gärdefors and Orrhage (1968)
Zooplankton	8	24-109	6	Plankton pump	Barnes and Marshall (1951) Cassie (1962)
Corn borer larvae	9	136-2,840	1	Individual corn plant	McGuire, Brindley, and Bancroft (1957)



**Figure 1.** A comparison of the fit of the (left) gamma and (right) Weibull distributions relative to the fit of the lognormal distribution, for the data described in Table 1 (fish,  $\circ$ ; polychaetes,  $\square$ ; ophiuroids,  $\triangle$ ; zooplankton,  $\blacksquare$ ; corn borer larvae,  $\diamond$ ). The diagonal line represents the 1 : 1 relationship. Symbols below represent cases where the gamma or Weibull distributions provided a better fit than the lognormal distributions. Maximum likelihood fits were made using the MLP package (Ross, 1987). Goodness of fit was measured using the likelihood ratio statistic,  $G^2$  (sometimes called the residual chi-squared).

**Table 2**

Number of samples that were positively skewed, after log-transformation, and number of samples in which samples were best fit by Weibull, gamma, and lognormal distributions. Number in parentheses represents number of samples that were significantly different ( $P < .05$ ) from Weibull, gamma, and lognormal distributions by using a goodness-of-fit test. The range in the number of degrees of freedom for the test is listed in parentheses in the second column.

Taxon	Number of samples (d.f.)	No. positively skewed after transformation	No. of samples best fit by (No. samples significantly different from)		
			Weibull distribution	Gamma distribution	Lognormal distribution
Fish	10 (3-9)	9	4 (2)	0 (2)	6 (1)
Polychaetes	49 (2-10)	48	21 (4)	10 (4)	18 (2)
Ophiuroids	2 (20-30)	0	1 (0)	1 (0)	0 (0)
Zooplankton	8 (2-13)	7	3 (2)	3 (1)	2 (2)
Corn borer larvae	9 (2-47)	7	5 (3)	3 (2)	1 (2)

**2. Estimators**

Consider a survey in which the density of organisms ( $x_i$  for  $i = 1, \dots, n$ ) is determined at  $n$  randomly chosen locations in sampling units of fixed size and in which the area surveyed is small relative to the area occupied by the population. The commonly used estimate of the mean density is the sample mean ( $\bar{x} = \sum_{i=1}^n x_i/n$ ) with an estimated variance of  $V_{est}(\bar{x}) = n^{-1} ([\sum_{i=1}^n (x_i - \bar{x})^2]/(n - 1))$ . If the log-transformed nonzero observations follow a normal distribution, improved estimators of the mean ( $\hat{\mu}_l$ ) and estimated variance of the estimated mean ( $V_{est}(\hat{\mu}_l)$ ) are possible. Pennington (1983) suggested the following minimum variance unbiased estimators:

$$\mu_l = \begin{cases} \frac{m}{n} e^{\bar{y}} G_m\left(\frac{1}{2} s^2\right) & \text{if } m > 1 \\ \frac{x_1}{n} & \text{if } m = 1 \\ 0 & \text{if } m = 0 \end{cases} \tag{1}$$

and

$$V_{est}(\mu_l) = \begin{cases} \frac{m}{n} e^{2\bar{y}} \left\{ \frac{m}{n} G_m^2\left(\frac{1}{2} s^2\right) - \frac{m-1}{n-1} G_m\left(\frac{m-2}{m-1} s^2\right) \right\} & \text{if } m > 1 \\ \left(\frac{x_1}{n}\right)^2 & \text{if } m = 1 \\ 0 & \text{if } m = 0 \end{cases} \tag{2}$$

where  $\bar{y}$  and  $s^2$  are the sample mean and variance, respectively, of the log of the nonzero values,  $m$  is the number of nonzero values,  $n$  is the total number of observations, and where  $G_m$  is

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

For cases with no zeros, the formula for the mean reduces to that given by Finney (1941).

**3. Simulations**

To investigate the robustness of the lognormal-based estimators, 25,000 samples of size  $q$  were generated using NAG (1987) subroutines, each comprising  $(q - p)$  observations from a lognormal and  $p$  observations from one of three alternative distributions. Pseudorandom numbers from a lognormal distribution were contaminated with pseudorandom numbers from gamma and Weibull distributions with the same mean and variance as the lognormal distribution. Similar contamination was introduced from an alternative transformation of the unit normal,

$$\theta[u\sigma + \sqrt{(u\sigma)^2 + 1}]^2, \tag{4}$$

where  $\theta$  and  $\sigma$  are positive parameters and  $u$  is a unit normal random variable (Johnson and Kotz, 1970, p. 268). These distributions were chosen because they have similar shapes to the lognormal distribution over much of their parameter ranges but they are less skewed. Samples sizes ( $q$ ) of 20 and 40 were investigated. The results depend only on the coefficient of variation rather than the absolute value of the mean and variance. We chose a mean of 10 and allowed the variance of the distribution from which the pseudorandom numbers

were generated to range from 1 to 900. This varied the coefficient of variation from .1 to 3. For coefficient of variation greater than 1, the shapes of the gamma and Weibull distributions change such that the density of  $x$  tends to infinity as  $x$  tends to zero.

The following four statistics were calculated for each mean, variance, and proportion of contamination tested:

$$\begin{aligned} \text{Relative bias of } \hat{\mu}_l &= \frac{E[\hat{\mu}_l - \mu]}{\mu} \\ \text{Efficiency of } \hat{\mu}_l &= \frac{E[\hat{\mu}_l - \mu]^2}{E[\bar{x} - \mu]^2} \\ \text{Relative bias of } V_{\text{est}}(\hat{\mu}_l) &= \frac{E[V_{\text{est}}(\hat{\mu}_l) - V(\hat{\mu}_l)]}{V(\hat{\mu}_l)} \\ \text{Efficiency of } V_{\text{est}}(\hat{\mu}_l) &= \frac{E[V_{\text{est}}(\hat{\mu}_l) - V(\hat{\mu}_l)]^2}{E[V_{\text{est}}(\bar{x}) - V(\bar{x})]^2} \end{aligned} \quad (5)$$

where the expectations  $V(\hat{\mu}_l)$  and  $V(\bar{x})$  were calculated over the 25,000 pseudorandom samples for each combination of simulation parameters. The efficiency is the mean squared error of the lognormal-based estimators relative to the mean squared error of the sample mean and variance.

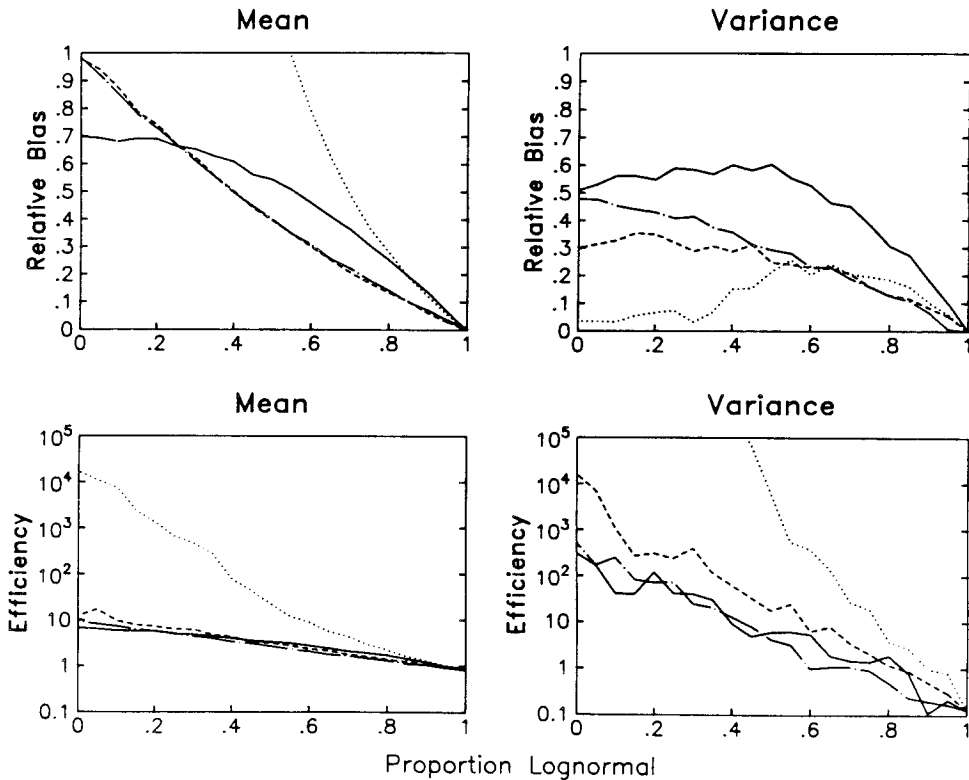
For each pseudorandom sample we determined the probability of rejecting the hypothesis that the sample was drawn from a lognormal distribution using the Shapiro and Wilk (1965) test. If the hypothesis was rejected, the relative bias and efficiency of the estimated mean and estimated variance for that simulation were not used in the cumulative analysis.

For our simulations, the function  $G_m(t)$  was estimated by summing the expression until the incremental change was less than .000001.

#### 4. Results

The results for coefficient of variation equal to 2 for the three contaminating distributions are displayed in Figure 2. A coefficient of variation of 2 is typical of many biological samples. The results for varying the coefficient of variation with contaminating data from a Weibull distribution are shown in Figure 3. The probability of not rejecting a sample ( $q = 20$ ) as being lognormal for Weibull contaminated data is given in Figure 4. The results can be summarized as follows:

1. The improved efficiency of the estimated mean and variance occurs only if there is little contamination—from 5% to 20% depending on the contaminating distribution.
2. Contamination can lead to large biases; e.g., for a coefficient of variation of 2, a 20% contamination results in a 15% to 30% bias in the estimated mean and variance (Fig. 2). If the distribution is gamma instead of lognormal the bias in the estimated mean can be greater than 200%.
3. The bias and loss of efficiency are relatively unaffected by the sample size (Fig. 2). That is, although the probability of rejecting contaminated data increases with increased sample size, the bias and loss of efficiency once a sample passes a test are only slightly reduced.
4. For small coefficient of variation, less than .30, both estimators have little or no bias and similar efficiencies.



**Figure 2.** The relative bias and the efficiency of the estimated mean and variance using lognormal-based estimators as a function of the proportion of data from a lognormal distribution. Sample sizes of 20 were used for contamination from a Weibull (dashed line), gamma (dotted line), and the alternative transformation of the normal described in the text (solid line). Contamination from the Weibull distribution with a sample size of 40 (long-short dashed line) is also shown. The coefficient of variation of all distributions was 2.

5. If there is not contamination, then the lognormal-based estimators yield increased efficiencies. Although the improvement for the mean is small, the increase in efficiency for the estimated variance can be large.
6. The probability of rejecting contaminated data is not large for small sample sizes (Fig. 4). The probability of rejection is largest at 50% to 75% contamination if the coefficient of variation is greater than 1 because the samples are often bimodal.

In the absence of contamination, the efficiency of lognormal-based estimators is directly related to sample size (Mehran, 1973; Smith, 1988) and inversely related to the proportion of zero observations in the data (Aitchison and Brown, 1957, p. 98). We carried out a series of simulations aimed at evaluating the accuracy of these conclusions. The proportion of zero observations was generated from pseudorandom samples from a binomial distribution. The simulation results were consistent with the earlier studies. In the presence of contamination, the results were qualitatively similar to Figures 2 and 3. We do not present figures of these results for the sake of brevity. However, the lognormal-based estimators become less efficient than the sample mean and variance at smaller levels of contamination as the sample size decreases or as the proportion of zero values in the data increases.

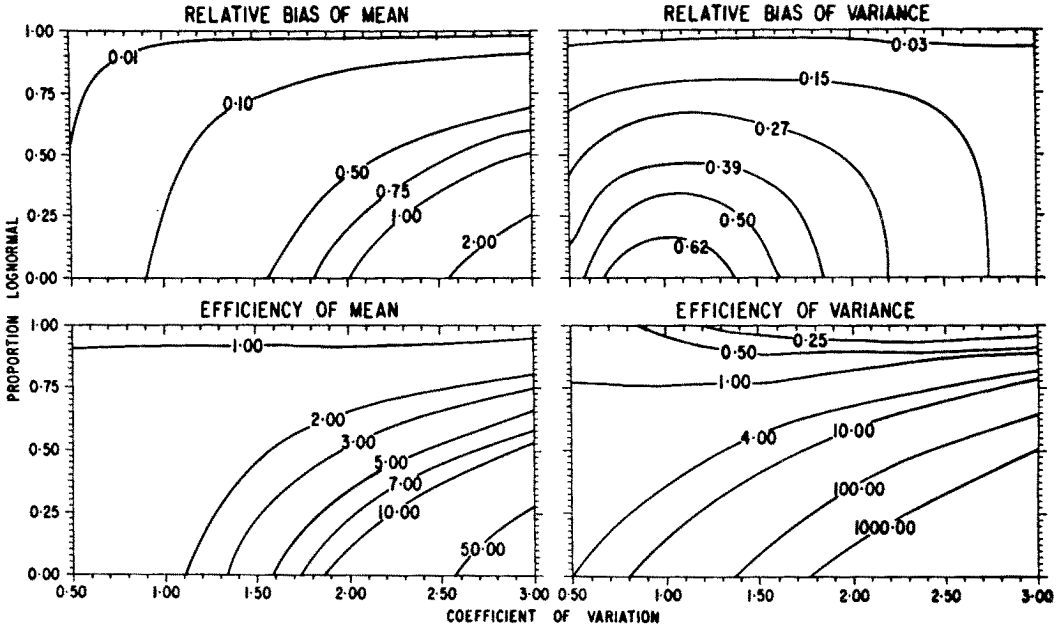


Figure 3. Contours of the relative bias and the efficiency of the estimated mean and variance using lognormal-based estimators as a function of the coefficient of variation of the pseudorandom numbers ( $x$ -axes) and the proportion of data from a lognormal distribution ( $y$ -axes) with contaminating data from a Weibull distribution.

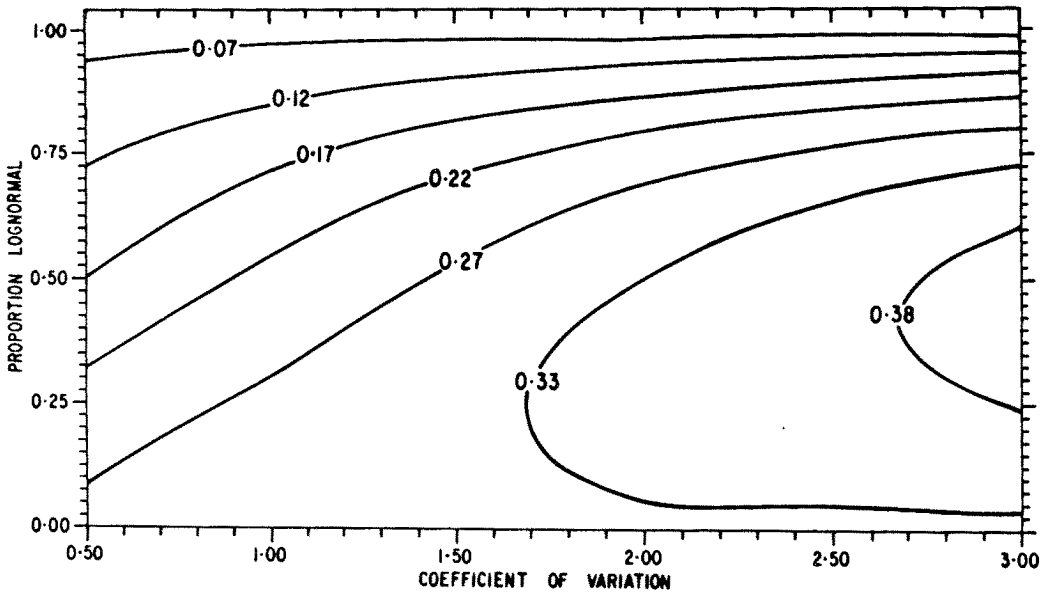


Figure 4. Contours for the probability of rejecting the hypothesis that data are lognormally distributed in cases when observations are drawn from a lognormal distribution contaminated with observations from a Weibull distribution. Sample size is 20. Axes are described in the legend for Figure 3.

## 5. Discussion

Lognormal-based estimators are very sensitive to violations of model assumptions. It must be clearly demonstrated that nonzero values follow a lognormal distribution consistently for lognormal-based estimators to be used accurately and with confidence. Deviations from the model assumptions, which greatly reduce the efficiency of the lognormal-based estimators, have a low probability of being detected for small sample sizes ( $\leq 40$ ). Typically, surveys of animal abundance (e.g., fish trawl surveys, plankton surveys) have fewer than 40 samples per stratum. We suggest that lognormal-based estimators of abundance be used only in cases where the adequacy of the lognormal assumption can be assured (e.g., if repeated samples from the same population consistently show the distribution to be lognormal).

Although we do not claim that the sample mean is a robust statistic (Huber, 1981), it is apparent that it provides a more robust estimator of abundance than the lognormal-based estimators, which are very sensitive to violations in model assumptions. Under these circumstances, the sample mean is the best estimator. Robust estimators designed to maintain Fisher consistency have not been developed for discrete skewed distributions and therefore present an area for future research (but see Kimber, 1983).

### ACKNOWLEDGEMENTS

We thank Nick Payton for his programming assistance. J. Hoenig, J. Rice, and S. Smith provided helpful comments.

### RÉSUMÉ

On teste la robustesse d'une méthode d'estimation de l'abondance qui suppose que la distribution des observations non nulles est lognormale (Pennington, 1983, *Biometrics* **39**, 281–286). Les écarts à cette hypothèse, qui ne peuvent pas être détectés à partir d'échantillons de taille modeste ( $\leq 40$ ), conduisent à des biais et à une diminution de l'efficacité. A moins de pouvoir montrer clairement, à partir d'échantillons répétés, que les valeurs non nulles suivent une distribution lognormale, la moyenne et la variance d'échantillonnage sont plus robustes que les estimateurs de la moyenne et de la variance de la population basés sur la lognormalité.

### REFERENCES

- Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution with Special Reference to Its Uses in Economics*. Cambridge: Cambridge University Press.
- Barnes, H. and Marshall, S. M. (1951). On the variability of replicate plankton samples and some applications of contagious series to the statistical distribution of catches over restricted periods. *The Journal of Marine Biological Association U.K.* **30**, 233–263.
- Brodie, W. B. and Wells, R. (1985). The distribution of trawl catches of cod and American plaice from research vessel surveys in NAFO divisions 3L, 3M and 3N. SCR Doc. 85/106. Halifax: North Atlantic Fisheries Organization.
- Cassie, R. M. (1962). Frequency distribution models in the ecology of plankton and other organisms. *Journal of Animal Ecology* **31**, 65–92.
- Finney, D. J. (1941). On the distribution of a variate whose logarithm is normally distributed. *Journal of the Royal Statistical Society, Supplement* **7**, 151.
- Gärdefors, D. and Orrhage, L. (1968). Patchiness of some marine bottom animals: A methodological study. *Oikos* **19**, 311–321.
- Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.
- Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions, Vol. 2*. New York: Wiley.
- Kimber, A. C. (1983). Trimming in gamma samples. *Applied Statistics* **32**, 7–14.
- Kirkegaard, J. B. (1969). A quantitative investigation of the central North Sea polychaeta. *Spolia Zoologica Musei Hauniensi* **29**, 1–285.

- McGuire, J. U., Brindley, T. A., and Bancroft, T. A. (1957). The distribution of European corn borer larvae *Pyrausta nubilalis* (HBN.), in field corn. *Biometrics* **13**, 65-78.
- Mehran, F. (1973). Variance of the MVUE for the lognormal mean. *Journal of the American Statistical Association* **68**, 726-727.
- Myers, R. A. and Pepin, P. (1986). The estimation of population size from research surveys using regression models C. M. 1986/D:9. Copenhagen: International Council for the Exploration of the Sea.
- NAG. (1987). *NAG Fortran Library Manual Mark 13*. Downers Grove, Illinois: Numerical Algorithms Group, Inc.
- Pennington, M. (1983). Efficient estimators of abundance, for fish and plankton surveys. *Biometrics* **39**, 281-286.
- Pennington, M. and Berrien, P. (1984). Measuring the precision of estimates of total egg production based on plankton surveys. *Journal of Plankton Research* **6**, 868-879.
- Reid, D. D. (1981). The Poisson lognormal distribution and its use as a model of plankton aggregation. In *Statistical Distributions in Scientific Work, Vol. 6*, C. Taillie et al. (eds), 303-316. New York: Plenum Press.
- Ross, G. J. S. (1987). *Maximum Likelihood Program*. Rothamsted: The Numerical Algorithms Group Limited.
- Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591-611.
- Sherman, K., Smith, W., Morse, W., Berman, M., Green, J., and Ejsymont, L. (1984). Spawning strategies of fishes in relation to circulation, phytoplankton production, and pulses in zooplankton off the northeastern United States. *Marine Ecology Progress Series* **18**, 1-19.
- Smith, S. J. (1988). Evaluating the efficiency of the  $\Delta$ -distribution mean estimator. *Biometrics* **44**, 485-493.
- Steinarsson, B. and Stefansson, G. (1986). Comparison of random and fixed trawl stations in Icelandic groundfish surveys and some computation considerations. C. M. 1986/D:13. Copenhagen: International Council for the Exploration of the Sea.
- Taylor, L. R. (1984). Assessing and interpreting the spatial distributions of insect populations. *Annual Review of Entomology* **29**, 321-357.

*Received March 1987; revised May 1988, April 1989, and January 1990; accepted May 1990.*