Estimating Mean Concentrations under Transformation for Environmental Data with Detection Limits

Author(s): R. H. Shumway, A. S. Azari and P. Johnson

Source: *Technometrics*, Aug., 1989, Vol. 31, No. 3 (Aug., 1989), pp. 347-356

Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality

Stable URL: http://www.jstor.com/stable/3556144

TECHNOMETRICS, AUGUST 1989, VOL. 31, NO. 3

Estimating Mean Concentrations Under Transformation for Environmental Data With Detection Limits

R. H. Shumway, A. S. Azari, and P. Johnson

Division of Statistics
University of California
Davis, CA 95616

The reporting procedures for potentially toxic pollutants are complicated by the fact that concentrations are measured using small samples that include a number of observations lying below some detection limit. Furthermore, there is often a small number of high concentrations observed in combination with a substantial number of low concentrations. This results in small, nonnormally distributed censored samples. This article presents maximum likelihood estimators for the mean of a population, based on censored samples that can be transformed to normality. The method estimates the optimal power transformation in the Box–Cox family by searching the censored-data likelihood. Maximum likelihood estimators for the mean in the transformed scale are calculated via the expectation-maximization algorithm. Estimates for the mean in the original scale are functions of the estimated mean and variance in the transformed population. Confidence intervals are computed using the delta method and the nonparametric percentile and bias-corrected percentile versions of Efron's bootstrap. A simulation study over sampling configurations expected with environmental data indicates that the delta method, combined with a reliable value for the power transformation, produces intervals with better coverage properties than the bootstrap intervals.

KEY WORDS: Bootstrap; Box–Cox transformation; Censored data; Delta method; EM algorithm; Maximum likelihood.

1. INTRODUCTION

The problems involved in monitoring levels of various substances in the environment for possible classification as toxic air contaminants are of public-health concern because of the possible carcinogenic effects that such substances might have on man. In general, dose is considered to be the best measure of carcinogenicity, and the mean concentration of a given substance can be used as a direct estimator for dose.

The true mean of the given substance is not known and must be estimated from a sample of observations collected from some environmental area of concern. Since measurements of toxic air contaminants involve potentially expensive sampling procedures and even more expensive laboratory determinations, the sample sizes used to characterize the mean level of a toxic contaminant are typically rather small. Even so, if the underlying data from which the sample has been drawn can be assumed to follow a normal distribution, the simple average of a sample of observations estimates the mean, and the t distribution yields the usual small-sample confidence interval.

It is unfortunate that, for most environmental data, the assumption that the underlying distribution is normal will not be appropriate, so the usual sample mean will not be a good estimator for the population mean. One reason is that measurements made on toxic pollutants tend to have more extreme observations than one would expect under normality, both at very high levels and at very low levels. The analysis of nonnormal data can be approached by attempting to find a transformation that produces a normally distributed population when applied to the data sample. The logarithmic transformation is often applied to pollution data, leading to an estimation procedure based on the lognormal distribution (see Blandford and Shumway 1982; Gilliom and Helsel 1986; Ringdal 1975). In this article, we will consider the more general class of power transformations due to Box and Cox (1964) [see Hinkley and Runger (1984) for a more recent discussion]. Even so, although the estimation of the mean in the new scale can then be approached in the usual way using normal theory, a method should be sought for transforming the estimators for the mean in the transformed scale back to an estimator for the mean in the original scale (see Carroll and Ruppert 1984; Rubin 1984).

A further complication is introduced by the fact that laboratory determinations can only be measured above some detection limit, making it impossible to know the exact sample values of observations collected that have apparent concentrations below this limit. Even if the data are normally distributed, this censoring makes it impossible to calculate the conventional sample mean since one will be uncertain as to what to use for the censored values. The problem of estimating the mean for censored data using the method of maximum likelihood (which yields the ordinary sample mean in the uncensored case) was first considered by Cohen (1959). Applications of parametric maximum likelihood to censored data problems can be found in the works of Ringdal (1975), Aitkin (1981), Blandford and Shumway (1982), Poirier (1976), Gilliom and Helsel (1986), Cohen (1976), and Cohn (1988).

To summarize, in this article we seek a method for estimating the mean using small samples of nonnormal environmental data that are subject to censoring because of lower detection limits. We will also develop procedures for computing confidence limits for the mean using large-sample theory for maximum likelihood estimators (MLE's) [via the delta method (Cramer 1946) and the nonparametric bootstrap (Efron 1979, 1982)]. We describe techniques in Sections 2, 3, and 4. An evaluation over a range of example data is presented in Section 5.

2. THE USE OF TRANSFORMATIONS

When environmental data cannot be modeled in terms of the normal distribution, the sample mean is not, in general, the best estimator for the population mean. In such cases it may be appropriate to search for a function of the data values that conforms more closely to the normal distribution. Although the use of the logarithmic transformation for purposes of stabilizing variances and transforming to normality is well established in the literature, there may be occasions in which other transformations or perhaps even no transformation at all may be more appropriate.

A more general class of power transformations due to Box and Cox (1964) (see Johnson and Wichern 1988, pp. 155–162) can be applied to the data and includes the logarithm, no transformation, and various other power laws as special cases. To set the notation, assume that N independent observations are available of which n_1 are observed and n_2 are below detection limits. Suppose that the sample is denoted by $x_1, x_2, \ldots, x_N$ and that, if the observation x_i is censored, we know only that $x_i \leq T_i$, where T_i is some lower detection threshold, allowed to differ for each sample value. If x_i is observed, we write $x_i > T_i$, although the value of T_i in this case is irrelevant.

Define the transformed variables

$$\begin{aligned} y_i &= (x_i^\lambda - 1)/\lambda, \qquad && \lambda \neq 0 \\ &= \ln x_i, && \lambda = 0, \end{aligned} \tag{1}$$

when x_i is observed and the transformed thresholds

$$\begin{aligned} T_i^* &= (T_i^\lambda - 1)/\lambda, \qquad && \lambda \neq 0 \\ &= \ln T_i, && \lambda = 0, \end{aligned} \tag{2}$$

when x_i is censored. This is the Box–Cox transformation (see Box and Cox 1964), which we consider for x_i positive ($y_i > -1/\lambda$) and $\lambda = 0, \frac{1}{4}, \frac{1}{2}$, and 1. Since environmental data are nonnegative, the power transformations considered in this article are limited to (a) no transformation ($\lambda = 1$), (b) fourth root ($\lambda = \frac{1}{4}$), (c) square root ($\lambda = \frac{1}{2}$), and (d) logarithmic ($\lambda = 0$). Using one of the transformations in this family will hopefully lead to a set of transformed data that approximately follows a normal distribution. There are two problems that occur when the use of a transformation is considered.

First, one must decide which of the power laws produces data that conform best to the normal distribution. Box and Cox (1964) proposed evaluating the likelihood function under each of the proposed transformations, assuming that the likelihood function of the transformed data is that of a set of normally distributed observations. Then one simply chooses the transformation that produces the largest value for the likelihood function. This procedure has the virtue that it can be done automatically using a computer, but this does not necessarily mean that it is the best method for choosing a transformation; a modification of a standard goodness-of-fit criterion such as the Shapiro–Wilk statistic might do even better (e.g., see Johnson and Wichern 1988, p. 150). It is also important to establish for any given type of experimental data some common transformation that produces normally distributed observations in the majority of cases. Later on, we shall see how the confidence intervals computed are affected by choosing the wrong transformation.

The second problem that one has when carrying out an analysis on transformed data is that the mean of the transformed data is rarely the parameter of interest. Although the transformed scale can be of great intererst (e.g., the use of the logarithmic scale in measuring seismic magnitudes) in some fields, the primary objective here is to obtain an estimate of the mean and a confidence interval in the original scale of measurement. The theoretical mean in the original scale will be a nonlinear function of the means and variances in the transformed scale. For $\lambda > 0$ in (1), we have

$$E(X) = \int_{-1/\lambda}^{\infty} (\lambda y + 1)^{1/\lambda} \varphi\left(\frac{y - \mu}{\sigma}\right) \frac{dy}{\sigma} \tag{3}$$

with

$$\varphi(x) = (2\pi)^{-1/2}\exp\{-\tfrac{1}{2}x^2\} \quad (4)$$

so that $E(X)$ is a nonlinear function of the parameters λ, μ, and σ. When $\lambda = 0$, corresponding to the lognormal distribution, we have

$$E(x) = \exp\{\mu + \tfrac{1}{2}\sigma^2\}. \quad (5)$$

The approximate values of $E(X)$ using $-\infty$ for the lower limit in (3) for $\lambda = \frac{1}{4}, \frac{1}{2}$, and 1 are shown in Table A.1 in the Appendix. Now, the likelihood function can be maximized over means and variances in the transformed scale using normal distribution theory. Since MLE's of functions are the corresponding functions of the MLE's, we easily obtain the MLE's for the mean in the original scale by simply taking the appropriate functions of the estimators in the transformed scale.

3. ESTIMATION OF THE MEAN

We continue the discussion begun in the last section relating to estimating the mean, first in the transformed scale, taking into account the censoring, and then in the original scale. Various procedures can be considered for estimating the mean of a set of observations when some of the observations are censored. For example, a simulation study using small normal data sets was performed by Gleit (1985), who compared the performance of the standard mean with several fill-in options involving the proxies zero, the threshold, and the conditional expectation given the censoring. He concluded that the procedure of filling in conditional expectations leads to an estimator that has the smallest mean squared error of those considered.

The approach taken in this article will be to use the maximum likelihood procedure to estimate the mean and variance parameters in the transformed scale under censoring and then to use the properties of MLE's to get the MLE's in the original scale. One problem with this approach is that the likelihood function for the censored data involves the cumulative normal distribution function, which produces a nonlinear likelihood that cannot be solved directly for the estimators in the transformed scale. An early treatment of this problem using maximum likelihood is that of Poirier (1976). A correction for the bias in the MLE in the case of lognormally distributed censored data was proposed by Cohn (1988).

Now we may write the log-likelihood of the original observations $x_1, \ldots, x_N$, assuming that the transformed observation $y_1, \ldots, y_N$ is normally (truncated) distributed with mean μ and variance σ^2, as

$$\ln L(\lambda, \mu, \sigma) = -\frac{n_1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{x_i > T_i}(y_i - \mu)^2 + \sum_{x_i > T_i} J_i(\lambda) + \sum_{x_i \le T_i}\ln\Phi(Z_i), \quad (6)$$

where the notation $x_i \le T_i$ denotes summing over the censored values, n_1 denotes the number observed, $J_i(\lambda)$ is the Jacobian

$$\begin{aligned} J_i(\lambda) &= (\lambda - 1)\ln x_i, & \lambda &\ne 0 \\ &= -\ln x_i, & \lambda &= 0, \end{aligned} \quad (7)$$

$$Z_i = (T_i^* - \mu)/\sigma, \quad (8)$$

and

$$\Phi(z) = \int_{-\infty}^{z}\varphi(x)\,dx \quad (9)$$

denotes the standard normal cdf.

Complicated likelihood functions involving missing or incompletely observed data can be maximized using the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). Aitkin (1981) and Blandford and Shumway (1982) applied this algorithm to the regression case. Basically, the approach makes use of the likelihood of the complete data under the assumption that the censored data points have been observed. One can then calculate the conditional expectation of this likelihood given the pattern of censoring. The EM algorithm guarantees that iteratively maximizing this restricted likelihood with respect to the mean and variance leads to a sequence of estimators that always increase the likelihood function (6) of the original censored data. Furthermore, for any fixed value of the transformation parameter, the results of Wu (1983) showed that the sequence converges, in this case, to local maximizers of the likelihood function.

The EM algorithm operates on the so-called complete-data log-likelihood

$$\ln L'(\lambda, \mu, \sigma) = -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2 + \sum_{i=1}^{N} J_i(\lambda), \quad (10)$$

written as though no observations were censored. Iterations are defined as successively maximizing the expectation of the complete-data log-likelihood (10) conditioned on the censoring pattern. For example, if μ_k and σ_k are the current estimators at iteration k, the EM algorithm obtains μ_{k+1} and σ_{k+1} by maximizing (E_k denotes the expectation over the normal density determined by μ_k and σ_k)

$$E_k[\ln L'(\lambda, \mu, \sigma) \mid \text{censoring}]$$

$$= -\frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{x_i > T_i} (y_i - \mu)^2$$

$$- \frac{1}{2\sigma^2} \sum_{x_i \le T_i} E_k[(Y_i - \mu)^2 \mid Y_i \le T_i^*]$$

$$+ \sum_{i=1}^{N} J_i(\lambda) \tag{11}$$

over μ and σ for a fixed λ. Scanning the resulting maximizers over λ leads to the final estimator. The updated values for μ and σ at each stage are computed using

$$\hat{\mu}_{k+1} = \frac{1}{N} \left\{ \sum_{x_i > T_i} y_i + \sum_{x_i \le T_i} E_k(Y_i \mid Y_i \le T_i^*) \right\} \tag{12}$$

and

$$\hat{\sigma}^2_{k+1} = \frac{1}{N} \left\{ \sum_{x_i > T_i} (y_i - \mu_k)^2 + \sum_{x_i \le T_i} E_k[(Y_i - \mu_k)^2 \mid Y_i \le T_i^*] \right\}, \tag{13}$$

where the conditional means and variances are computed from

$$E_k[Y_i \mid Y_i \le T_i^*] = \mu_k - \sigma_k \frac{\varphi(Z_i)}{\Phi(Z_i)} \tag{14}$$

and

$$E_k[(Y_i - \mu)^2 \mid Y_i \le T_i^*] = \sigma_k^2 \left(1 - Z_i \frac{\varphi(Z_i)}{\Phi(Z_i)}\right), \tag{15}$$

where Z_i is defined as (8) evaluated at μ_k, σ_k. Note that the conditional expectation (14) is just that recommended as the best "fill-in" option by Gleit (1985). The preceding procedure leads to MLE's for λ, μ, and σ and hence, by substituting, into (3) or (5) for $E(X)$.

4. CONFIDENCE INTERVALS

The procedure described in the previous section leads to MLE's for the means in the original scale, but it does not immediately produce an estimator for the variance or a confidence interval. With a confidence interval, one can make an assessment of the probable range within which the true mean can be expected to lie.

In developing confidence intervals, we are faced with choices. Note that the MLE's developed previously will have a limiting normal distribution with a predictable mean and variance in large samples. This leads to the use of the delta method, usually credited to Cramer (1946). If we are not comfortable with large-sample theory because our samples are generally not large, then resampling via the bootstrap (see Efron 1979, 1981, 1982, 1985) may be useful. The adaptation of these two methods to the problems at hand will be described.

4.1 The Delta Method

The computation of the large-sample variance–covariance matrix of the mean and variance in the transformed scale depends on technical manipulations to compute the second derivatives of the log-likelihood function. The computation of the large-sample variance of the mean in the original scale depends on expanding it in a Taylor series about the mean and variance in the transformed scale and then using the central limit result of Cramer (1946) for functions of asymptotically normal variables.

To illustrate, a confidence interval for $E(X)$ follows that is a function of the parameter vector $\boldsymbol{\theta} = (\mu, \sigma)'$. It is convenient and realistic to regard λ as being fixed after the arguments of Hinkley and Runger (1984) and the accompanying discussions. An estimator for the variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ in this case (see also Poirier 1976) is

$$\hat{\text{cov}}(\hat{\boldsymbol{\theta}}) = \left\{ -\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\}_{\hat{\boldsymbol{\theta}}}^{-1}, \tag{16}$$

where $\ln L$ is abbreviated for (6). The elements of

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} L\mu\mu & L\mu\sigma \\ L\sigma\mu & L\sigma\sigma \end{pmatrix} \tag{17}$$

are given in the Appendix. The basis for the delta method of Cramer (1946) is that $\hat{E}(X)$ is a function of the elements of $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})'$, which are expected to be jointly asymptotically normal by the arguments of the Appendix so that $\text{var}(\hat{E}(X))$ can be estimated by

$$\hat{\text{var}}(\hat{E}(X)) = \left(\frac{\partial E(X)}{\partial \boldsymbol{\theta}}\right)_{\hat{\boldsymbol{\theta}}}' \hat{\text{cov}}(\hat{\boldsymbol{\theta}}) \left(\frac{\partial E(X)}{\partial \boldsymbol{\theta}}\right)_{\hat{\boldsymbol{\theta}}}. \tag{18}$$

The 2×1 vectors of partial derivatives of $E(X)$ are as given in Table A.1 in the Appendix. This implies an approximate $100(1 - \alpha)\%$ confidence interval of the form

$$\hat{E}(X) \pm Z_{\alpha/2} \sqrt{\hat{\text{var}}(\hat{E}(X))}, \tag{19}$$

where $Z_{\alpha/2}$ is the upper $100(1 - \alpha/2)$ percentile of the standard normal.

4.2 Bootstrap Methods

The nonparametric bootstrap of Efron (1979) is a resampling method that develops confidence intervals from the sampling distribution of the resampled data. To illustrate, suppose that we have a sample of N observations on a toxic contaminant and that several of the observations have only been measured as being below some fixed detection limit. The method of Section 3 is used to compute an MLE for the mean $E(X)$ in the original scale, and we wish to have an idea of the basic uncertainty in this estimator.

Consider drawing a sample of N observations with replacement from the original sample in the previous paragraph. Since the sample is drawn with replacement so that the same element can appear more than once, the sample drawn will likely be different from the original sample and will yield a different estimator for the mean. This is called a bootstrap sample. This procedure can be repeated a large number of times, say B, and each time one will, with high probability, obtain a different sample and a different estimator for the mean, say $\hat{E}_i(X)$ $(i = 1, 2, \ldots, B)$. All of these MLE's for the mean can be combined into an empirical cdf $\hat{G}(\cdot)$ by arranging them in ascending order from smallest to largest. Within this sample, one can compute for any given α the percentiles $\hat{G}^{-1}(\alpha/2)$ and $\hat{G}^{-1}(1 - \alpha/2)$, which are such that $100(\alpha/2)$ and $100(1 - \alpha/2)\%$ of the estimated means lie, respectively, below or at the values. The resulting lower and upper values in this case define the $100(1 - \alpha)\%$ bootstrap confidence interval. Efron (1981) showed that this iid resampling for censored data is preferable over more complicated methods. Efron (1985) also proposed a procedure leading to a bias-corrected (BC) bootstrap interval that adjusts the bootstrap percentile distribution according to where the MLE appears on that distribution. In particular, let $z_0 = \Phi^{-1}[\hat{G}(\hat{E}(X))]$ be the standard normal value corresponding to the MLE $\hat{E}(x)$. The BC $100(1 - \alpha)\%$ confidence interval is taken as the interval $\hat{G}^{-1}[\Phi(2z_0 - Z_{\alpha/2})]$ to $\hat{G}^{-1}[\Phi(2z_0 + Z_{\alpha/2})]$, where $Z_{\alpha/2}$ is the upper $100(1 - \alpha/2)$ percentile of the standard normal distribution. A review of bootstrap methodology was given by Efron and Tibshirani (1986).

5. SIMULATIONS AND AN EXAMPLE

To be able to choose a method to recommend from those presented in Sections 2–4, a simulation study was designed using the following files as inputs: (a) 1,000 samples of size 20 with 10% censoring, (b) 1,000 samples of size 20 with 20% censoring, (c) 1,000 samples of size 50 with 10% censoring, and (d) 1,000 samples of size 50 with 20% censoring. The preceding samples were generated for (a) normally distributed data, (b) data for which the square root was normal, and (c) lognormally distributed data, making a total of 12 files. To ensure comparability, all files were started with the same random number seed. The normal distribution chosen for the untransformed data had $\mu = 3$ and $\sigma = 1$. The sample size chosen ensures that the 90% coverage probabilities will be estimated to within $\pm.02$ with 95% confidence.

The method used for estimating $E(X)$ first transforms the variables using (1) and (2) for a particular $\lambda(0, \frac{1}{4}, \frac{1}{2}, 1)$ and computes the mean and sample standard deviations for the transformed variables as initial estimators for μ and σ. Successive updates for μ and σ, say from μ_k, σ_k to μ_{k+1}, σ_{k+1}, are defined using Equations (5) and (6). At each stage, the log-likelihood (6), say L_k, is monitored, and iterations terminate when $|(L_{k+1} - L_k)/L_k| < 10^{-6}$ or when the number of iterations exceeds 25. Table 1 shows a typical set of iterates for a lognormal sample of size 20 with 10% censoring. The sequence is typical even when the λ is not the correct one. In general, cases in which the procedure failed to converge in 25 or fewer iterations were extremely rare. (In 1,000 bootstrap samples, this phenomenon never occurred more than twice.)

Table 2 shows how well the Box–Cox method for choosing the best power transformation performed on the 1,000 samples. Note that the method works quite well when choosing no transformation (correct about 70% of the time) or in choosing the logarithmic transformation (correct 48%–68% of the time). Note that in the case of the logarithmic transformation, the method worked substantially better for the large samples. The differences between 10% and 20% censoring were minor. The method has some difficulty choosing the square-root transformation when it has been applied; this correct decision is made in only about half the cases.

Table 3 shows the estimated coverages of confidence intervals computed by three methods when the initial choice for the power transformation is as determined in Table 2. Since the nominal level for these intervals was chosen to be 90%, one can evaluate how well the intervals are covering the known true value by comparing the tabular entry to 900. Of course, the expected lengths (averaged over 1,000 samples) of the intervals, given below the proportion, are important as well; given that the coverages

Table 1. A Typical Sequence of Iterates Using the EM Algorithm on a Censored Sample of Size N = 20

Iteration k	Mean μ_k	Standard deviation σ_k	Log-likelihood L_k
1	2.874	.984	−65.55568
2	2.777	1.107	−65.07636
3	2.756	1.140	−65.04903
4	2.750	1.150	−65.04673
5	2.748	1.153	−65.04652
6	2.748	1.154	−65.04652

Table 2. Simulation Results for Censored Mean Estimation Using the Box–Cox Transformation Showing the Proportion of Times Each Transformation Was Chosen (1,000 samples)

			Transformation chosen			
Transformation	Sample size	Censoring	None	Square root	Fourth root	Log
None	20	10%	.66	.14	.05	.15
	20	20%	.69	.09	.05	.17
	50	10%	.71	.15	.07	.07
	50	20%	.69	.13	.07	.11
Square root	20	10%	.45	.20	.14	.21
	20	20%	.52	.16	.10	.22
	50	10%	.33	.34	.18	.15
	50	20%	.39	.29	.14	.18
Log	20	10%	.02	.13	.28	.57
	20	20%	.07	.18	.25	.50
	50	10%	.00	.03	.25	.72
	50	20%	.00	.08	.26	.66

are equal, one would prefer the shorter interval. Note that the theoretical values of $E(X)$ were 4, 6.5, and 33.115 for the three transformations corresponding to $\lambda = 1$, .5, and 0, respectively, so the expected lengths over different λ are not comparable.

The delta method performed best overall with coverages of .89–.90 for the normal and square root normal data and .82–.87 for the lognormal data. The BC percentile interval (based on 1,000 bootstrap replications) was next, but it seemed to suffer some degradation as the censoring moved from 10% to 20%. The percentile intervals tend to miss on the high side, whereas the opposite is true for the BC percentiles. The delta method has roughly an equal number of mistakes on each side with the exception that the lognormal case has more intervals erring on the low side. The other methods did less well, particularly on the lognormal data. The relatively less successful performance of the bootstrap-based percentile methods is not a complete surprise; this was studied by Loh (1987). A small-scale simulation in Shumway and Azari (1988) gave some indication of the bias and mean squared error to be expected for the maximum likelihood procedure. They found that the average bias of the maximum likelihood estimates was less than 1% of the mean. Cohn (1988) proposed a correction procedure for reducing the bias in cases in which it cannot be neglected.

It is interesting that all methods did less well for the lognormal data when evaluated in terms of coverage probabilities. It was conjectured that the poorer coverage might have been caused by the situations in which the Box–Cox method recommended some-

Table 3. Simulation Results for the Three Methods for Censored Mean Estimation Using the Box–Cox Transformation (1,000 bootstrap replications, 1,000 samples, nominal 90% coverage)

			Delta method			Percentile			BC percentile		
Transformation	Sample size	Censoring	C	L	H	C	L	H	C	L	H
None	20	10%	899	47 (.72)	54	836	107 (.65)	57	880	40 (.66)	80
	20	20%	897	40 (.73)	63	767	170 (.56)	63	823	51 (.59)	126
	50	10%	901	38 (.46)	61	823	130 (.42)	47	869	31 (.42)	100
	50	20%	894	37 (.47)	69	744	201 (.36)	55	748	44 (.37)	69
Square root	20	10%	897	63 (1.85)	40	830	126 (1.68)	44	892	45 (1.73)	63
	20	20%	894	65 (1.91)	41	756	205 (1.50)	39	833	69 (1.59)	98
	50	10%	903	54 (1.19)	43	826	141 (1.08)	33	878	35 (1.11)	87
	50	20%	897	59 (1.21)	44	739	236 (.97)	25	796	49 (1.00)	155
Log	20	10%	829	161 (27.12)	10	763	221 (25.57)	16	843	108 (31.03)	49
	20	20%	822	166 (27.47)	12	697	290 (23.27)	13	797	111 (31.30)	92
	50	10%	896	118 (17.98)	13	784	206 (16.71)	10	847	57 (19.95)	96
	50	20%	858	127 (18.00)	15	652	342 (15.20)	6	724	71 (17.77)	205

NOTE: Average lengths are in parentheses. *C* denotes the number of intervals covering the true mean, *L* denotes the number missing on the low side, and *H* denotes the number missing on the high side.

thing other than the logarithmic transformation. There is strong support in the literature (e.g., see Hinkley and Runger 1984) for the idea of doing the analysis assuming a fixed power transformation. Table 4 shows the coverages resulting from fixing the transformation at no transformation or at the logarithmic and applying the results to normal and lognormal data. It is clear that applying the logarithmic transformation when the data are, in fact, lognormal brings the coverages right up to .88–.90. Hence it is probably the incorrect Box–Cox choices in the original simulation of Table 3 that are causing the reduced coverages. Applying no transformation when the data are lognormally distributed results in a severe reduction in coverage (.75–.83). Hence it is clear that estimating the transformation as in Table 3 yields better coverages (.82–.87); knowing the right transformation as in Table 4 does even better. The upper part of Table 4 is interesting in that we do not seem to pay much of a penalty in coverage for assuming that a lognormal transformation is appropriate for normally distributed data. It is plausible that varying λ over the bootstrap replications might improve the coverage. In a smaller-scale simulation involving 200 samples of lognormally distributed data ($N = 20$, 10% censoring), this approach gave essentially the same coverages as in Table 3 (.83, .78, .86). Note also that the bias of the estimators was not larger in the case in which the data were lognormally distributed and the λ was estimated using the Box–Cox method.

To summarize, the delta method seems to do the best over all alternatives. Since the coverage can be affected either by choosing the wrong transformation all of the time or by choosing the wrong transformation some of the time, it is probably advisable to run the Box–Cox method over a reasonable number of environmental samples and then use the transformation maximizing the likelihood to construct the confidence intervals for all samples.

The computing for this study was done in a variety of environments ranging from a PC-XT clone to a mainframe VAX. Generally, 1,000 bootstrap replications for a sample of size 20 will take about 10 minutes on a PC-XT equipped with an 8087 math computation chip or about one minute on a COMPAQ with an 80386 processor. The delta method by itself is very fast on any machine. Program documentation and listings for microcomputers and mainframe computers are available in Shumway and Azari (1988).

We illustrate the approach on a 24-hour field sample taken during ambient air monitoring of the potentially toxic air pollutant ethyl parathion at Heber Station in the Imperial Valley, California. The observed concentrations are .010(2), .010(2), .010(2), .010(2), .018(1), .032(1), .012(1), .015(1), .010(2), .078(1), .092(1), .023(1), .018(1), and .010(1). [Values in parentheses indicate observed (1) or censored (2).] Units are $\mu g/m^3$. These data show 5 of 14 values below the detection limits. Various means computed by the investigators for this sample gave the estimators .033 for samples positive for ethyl parathion only, .025 for the mean value replacing the censored values by .01, and .023 for the mean value replacing the censored values by .005. Applying the Box–Cox

Table 4. Simulation Showing the Result of Fixing the Transformation on Normal and Lognormal Samples (1,000 samples, nominal 90% coverage) and Using the Delta Method

			Normal assumed			Lognormal assumed		
Transformation	*Sample size*	*Censoring*	*C*	*L*	*H*	*C*	*L*	*H*
None	20	10%	899	47 (.73)	54	907	37 (.73)	54
	20	20%	899	40 (.75)	61	899	31 (.71)	70
	50	10%	900	42 (.47)	58	910	25 (.47)	65
	50	20%	899	38 (.48)	63	882	22 (.45)	96
Log	20	10%	802	186 (29.1)	12	883	111 (31.5)	6
	20	20%	781	215 (32.4)	4	883	113 (32.2)	4
	50	10%	834	158 (20.1)	8	904	.85 (19.5)	11
	50	20%	752	246 (22.1)	2	899	.89 (19.7)	12

NOTE: *C* denotes the number of intervals covering the true mean, *L* denotes the number missing on the low side, and *H* denotes the number missing on the high side.

Table 5. 90% Confidence Intervals for Ethyl Parathion Means, Imperial Valley, California

Method	Lower limit	Upper limit
Delta	.0101	.0366
Percentile	.0118	.0299
BC percentile	.0184	.0446

NOTE: The estimated mean is .0233.

procedure gave values for the log-likelihood of 24.58, 24.29, 23.73, and 21.90, corresponding to the logarithmic, fourth root, square root, and no transformation, respectively; accordingly, the log transformation was used in the analysis yielding the MLE .0233, which is closest to the mean substituting .005 for the censored values. The 90% confidence intervals are shown in Table 5 and indicate that the three methods give somewhat different results with the bootstrap method giving shorter intervals. On the basis of the simulations, these intervals probably will not attain a coverage of 90% in the long run.

6. REMARKS

This article has concentrated on the development of a transformation using maximum likelihood and then retransforming as a reasonable procedure for estimating the mean of a small sample of nonnormally distributed environmental data that has some observations below the detection limit.

In general, we concluded, through the use of simulations, that the confidence intervals for the mean were best when the correct transformation is known and less well determined when the transformation had to be estimated. Hence it seems best to analyze a large number of consistent samples with the hope that they all follow approximately the same probability law; for example, they may all be approximately lognormally distributed.

The simulations also indicated that, over the conditions considered—namely, sample sizes of 20 and 50 with 10% and 20% censoring—the approximate delta-method intervals based on the large-sample properties of the maximum likelihood did better than bootstrapping. If the transformation was known, the 90% delta-method intervals covered the true mean about 90% of the time in all cases, whereas the bootstrap coverages were as low as 62% for a 90% nominal interval. If the transformation was not known, the coverages achieved by the delta method were generally at least 85% for a nominal 90% interval. Hence the delta method is recommended on the basis of this study.

ACKNOWLEDGMENTS

This research was supported in part by California Air Resources Board Contract A733-045. John Moore of the Air Resources Board posed the problem considered in this report and contributed to the solution. We are grateful to P. K. Bhattacharya of the Division of Statistics, University of California, Davis, for discussions relating to the asymptotics of the Appendix. We are also indebted to the editor, an associate editor, and two insightful reviewers who contributed numerous suggestions, including a more detailed investigation of the delta method.

APPENDIX: ASYMPTOTIC DISTRIBUTION THEORY

We summarize first the equations for the mean function and derivatives for various values of λ. Table A.1 shows the mean functions, evaluated for $\lambda \neq 0$ by Equation (3). The derivatives needed for calculating the variance term (18) for the delta method are also shown.

To calculate the variance, we need expressions for the elements of the information matrix (17) which become

$$L\mu\mu = -\frac{n_1}{\sigma^2} - \frac{1}{\sigma^2} \sum_{x_i \le T_i} (Z_i R_i + R_i^2), \quad \text{(A.1)}$$

$$L\mu\sigma = -\frac{2}{\sigma^3} \sum_{x_i > T_i} (y_i - \mu) - \frac{1}{\sigma^2} \sum_{x_i \le T_i} (Z_i R_i^2 + Z_i^2 R_i - R_i), \quad \text{(A.2)}$$

and

$$L\sigma\sigma = \frac{n_1}{\sigma^2} - \frac{3}{\sigma^4} \sum_{x_i > T_i} (y_i - \mu)^2 - \frac{1}{\sigma^2} \sum_{x_i \le T_i} (Z_i^2 R_i^2 + Z_i^3 R_i - 2 Z_i R_i), \quad \text{(A.3)}$$

where

$$R_i = \frac{\varphi(Z_i)}{\Phi(Z_i)} \quad \text{(A.4)}$$

Table A.1. Mean Functions in the Original Scale and Partial Derivatives

λ	$E(x)$	$\partial E(x)/\partial\mu$	$\partial E(x)/\partial\sigma$
0	$\exp\{\mu + \frac{1}{2}\sigma^2\}$	$\exp\{\mu + \frac{1}{2}\sigma^2\}$	$\sigma \exp\{\mu + \frac{1}{2}\sigma^2\}$
.25	$\frac{3}{16}\sigma^2 + \frac{3}{8}\sigma^2(\frac{1}{4}\mu + 1)^2 + (\frac{1}{4}\mu + 1)^4$	$(\frac{1}{4}\mu + 1)^3 + \frac{3}{16}\sigma^2(\frac{1}{4}\mu + 1)$	$\frac{3}{8}\sigma + \frac{3}{4}(\frac{1}{4}\mu + 1)^2$
.50	$(\frac{1}{2}\mu + 1)^2 + \frac{1}{4}\sigma^2$	$(\frac{1}{2}\mu + 1)$	$\frac{1}{2}\sigma$
1.00	$(\mu + 1)$	1	0

(see also Cohen 1976). To justify using the normal approximation and the delta method in Section 4.1, we will need the corresponding versions of (14) and (15) for $y_i > T_i^*$, which are

$$E[Y_i \mid Y_i > T^*] = \mu + \sigma \frac{\varphi(Z_i)}{\Phi(-Z_i)} \quad (A.5)$$

and

$$E[(Y_i - \mu)^2 \mid Y_i > T^*] = \sigma^2 \left(1 + Z_i \frac{\varphi(Z_i)}{\Phi(-Z_i)}\right). \quad (A.6)$$

We sketch a central limit result for the asymptotic distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\mu, \sigma)'$ and $\hat{\boldsymbol{\theta}}$ is the MLE (see Cox and Hinkley 1974, sec. 9.2). First, note that the elements of the vector $\partial \ln L/\partial \boldsymbol{\theta} = (L_\mu, L_\sigma)'$ can be written as

$$L_\mu = \frac{1}{\sigma^2} \sum_{i=1}^{N} I_{[y_i > T_i^*]}(y_i - \mu) - \frac{1}{\sigma} \sum_{i=1}^{N} I_{[y_i \le T_i^*]} R_i \quad (A.7)$$

and

$$L_\sigma = \frac{1}{\sigma} \sum_{i=1}^{N} I_{[y_i > T_i^*]} + \frac{1}{\sigma^3} \sum_{i=1}^{N} I_{[y_i > T_i^*]}(y_i - \mu)^2 - \frac{1}{\sigma} \sum_{i=1}^{N} I_{[y_i \le T_i^*]} Z_i R_i, \quad (A.8)$$

where $I_{[y_i > T_i^*]}$ and $I_{[y_i \le T_i^*]}$ are the indicator functions of the observed and censored values and the argument Z_i is omitted from R_i. Note that L_μ and L_σ are both sums of non-iid zero-mean random variables and that $E(L_\mu)$ and $E(L_\sigma) = 0$ are both 0. For example, writing L_μ as ΣU_i, it follows that

$$E(U_i) = \frac{1}{\sigma^2}(1 - \Phi_i) \frac{\sigma \varphi_i}{1 - \Phi_i} - \frac{1}{\sigma} \Phi_i \frac{\varphi_i}{\Phi_i} = 0,$$

where the argument Z_i is again omitted from φ_i and Φ_i. Furthermore, rewriting (A.1)–(A.3) in terms of indicator functions and taking expectations leads to

$$E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = \begin{pmatrix} E(L\mu\mu) & E(L\mu\sigma) \\ E(L\sigma\mu) & E(L\sigma\sigma) \end{pmatrix}, \quad (A.9)$$

where

$$E(L_{\mu\mu}) = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \left(Z_i \varphi_i + 1 - \Phi_i + \frac{\varphi_i^2}{\Phi_i}\right), \quad (A.10)$$

$$E(L_{\mu\sigma}) = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \left(Z_i^2 \varphi_i + Z_i \frac{\varphi_i^2}{\Phi_i} + \varphi_i\right), \quad (A.11)$$

and

$$E(L_{\mu\sigma}) = -\frac{1}{\sigma^2} \times \sum_{i=1}^{N} \left(Z_i^3 \varphi_i + Z_i^2 \frac{\varphi_i^2}{\Phi_i} + Z_i \varphi_i + 1 - \Phi_i\right). \quad (A.12)$$

Dividing through by N implies that all terms in (A.10)–(A.12) are of the form

$$\frac{1}{N} \sum_{i=1}^{N} g(z_i) = \int_{-\infty}^{\infty} g(z)\, dF_N(z),$$

where $F_N(z) = 1/N^{-1} \#\{Z_i \le z\}$ is the empirical cdf of the scaled thresholds $Z_i = (T_i^* - \mu)/\sigma$ for $i = 1, \ldots, N$. Now assume that $F_N(z) \to F(z)$, the limiting distribution function of the Z_i's. Then, for $g(z)$ bounded and continuous,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} g(z_i) = \int_{-\infty}^{\infty} g(z)\, dF(z)$$

by the Helly–Bray theorem (see Rao 1973, p. 117). To show the limiting behavior of

$$A_N = -\frac{1}{N} E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right), \quad (A.13)$$

it only remains to show that all arguments of (A.10)–(A.12) are bounded and continuous; note that the only questionable assertion relates to the boundedness of $z^2\varphi^2(z)/\Phi(z)$, which follows from standard bounds for $\Phi(z)$, as in Feller (1957, p. 166). Hence it follows that (A.10)–(A.12) when divided by N converge to the appropriate integral expressions. For example,

$$\lim_{N\to\infty} (a_{11}^N) = -\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E(L_{\mu\mu}) = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left[z\varphi(z) + \Phi(-z) + \frac{\varphi^2(z)}{\Phi(z)}\right] dF(z). \quad (A.14)$$

Assume also that the limiting matrix $A = \lim_{N\to\infty} A_N$ is nonsingular.

We may also verify a central limit result for $N(L_\mu, L_\sigma)$, where L_μ and L_σ are as given in (A.7) and (A.8) using the preceding limiting arguments and the Lindeberg–Feller version of the central limit theorem (see Rao 1973, p. 128) combined with the multivariate central limit theorem. The regularity conditions on the third derivatives follow by differentiation of (A.1)–(A.3).

We summarize by stating that, subject to convergence of the empirical cdf's for the scaled thresholds and to the nonsingularity of the limiting matrix A, the asymptotic distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is normal with mean $\mathbf{0}$ and variance covariance matrix A^{-1}.

[*Received February 1988. Revised November 1988.*]

REFERENCES

Aitkin, M. (1981), "A Note on the Regression Analysis of Censored Data," *Technometrics*, 23, 161–163.

Blandford, R. R., and Shumway, R. H. (1982), "Magnitude-Yield for Nuclear Explosions With Stations Containing Clipped and Low Amplitude Signals," Technical Report VSC-12, Teledyne-Geotech, Seismic Data Analysis Center, Alexandria, Virginia.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 211–252.

Carroll, R. J., and Ruppert, D. (1984), Comment on "The Analysis of Transformed Data," by D. V. Hinkley and G. Runger, *Journal of the American Statistical Association*, 79, 312–313.

Cohen, A. C. (1959), "Simplified Estimators for the Normal Distribution When the Samples Are Singly Censored or Truncated," *Technometrics*, 1, 217–237.

——— (1976), "Progressively Censored Sampling in the Three Parameter Log-Normal Distribution," *Technometrics*, 18, 99–103.

Cohn, T. A. (1988), "Adjusted Maximum Likelihood Estimation of the Moments of Lognormal Populations From Type-I Censored Samples," Open-File Report 88-350, U.S. Geological Survey, Reston, Virginia.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Cramer, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312–319.

——— (1982), "The Jackknife, the Bootstrap and Other Resampling Plans," CBMS/NSF Monogram 38, Society for Industrial and Applied Mathematics, Philadelphia.

——— (1985), "Bootstrap Confidence Intervals for a Class of Parametric Problems," *Biometrika*, 72, 45–58.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54–77.

Feller, W. (1957), *Probability Theory and Its Applications* (Vol. 1), New York: John Wiley.

Gilliom, R. H., and Helsel, D. R. (1986), "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data 1 and 2," *Water Resources Research*, 22, 135–155.

Gleit, A. (1985), "Estimation for Small Normal Data Sets With Detection Limits," *Environmental Science and Technology*, 19, 1206–1213.

Hinkley, D. V., and Runger, G. (1984), "The Analysis of Transformed Data" (with discussion), *Journal of the American Statistical Association*, 79, 302–320.

Johnson, R. A., and Wichern, D. W. (1988), *Applied Multivariate Statistical Analysis* (2nd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Loh, W. (1987), "Calibrating Confidence Coefficients," *Journal of the American Statistical Association*, 82, 155–162.

Poirier, D. J. (1976), "The Use of the Box-Cox Transformation in Limited Dependent Variable Models," *Journal of the American Statistical Association*, 73, 284–287.

Rao, C. R. (1973), *Linear Statistical Inference and Its Application* (2nd ed.), New York: John Wiley.

Ringdal, F. (1975), "Maximum Likelihood of Seismic Magnitude," *Bulletin of the Seismological Society of America*, 66, 789–802.

Rubin, D. B. (1984), Comment on "The Analysis of Transformed Data," by D. V. Hinkley and G. Runger, *Journal of the American Statistical Association*, 79, 309–312.

Shumway, R. H., and Azari, A. S. (1988), "Estimating Mean Concentrations When Some Data Are Below the Detection Limit," final report, California Air Resources Board Contract A733-045.

Wu, C. F. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.