WATER RESOURCES BULLETIN
VOL. 25, NO. 4 AMERICAN WATER RESOURCES ASSOCIATION AUGUST 1989

ESTIMATING MEAN AND VARIANCE FOR ENVIRONMENTAL SAMPLES WITH BELOW DETECTION LIMIT OBSERVATIONS[1]

Michael C. Newman, Philip M. Dixon, Brian B. Looney, and John E. Pinder, III[2]

ABSTRACT: Left-censoring of data sets complicates subsequent statistical analyses. Generally, substitution or deletion methods provide poor estimates of the mean and variance of censored samples. These substitution and deletion methods include the use of values above the detection limit (DL) only, or substitution of 0, DL/2 or the DL for the below DL values during the calculation of mean and variance. A variety of statistical methods provides better estimators for different types of distributions and censoring. Maximum likelihood and order statistics methods compare favorably to the substitution or deletion methods. Selected statistical methods applicable to left-censoring of environmental data sets are reviewed with the purpose of demonstrating the use of these statistical methods for coping with Type I (and Type II) left-censoring of normally and log-normally distributed environmental data sets. A PC program (UNCENSOR) is presented that implements these statistical methods. Problems associated with data sets with multiple DLs are discussed relative to censoring methods for life and fatigue tests as recently applied to water quality data sets.
(KEY TERMS: mean; variance; detection limit; censored data; environmental data; statistical analysis.)

INTRODUCTION

The occurrence of values below the detection limit (DL) in environmental data sets is a major statistical complication. Uncertainty about the actual values below the DL can bias or preclude subsequent statistical analyses. Consequently, the use of such data (left-censored) for defining conditions and detecting trends or relationships can be compromised despite the most rigorous quality assurance program.

The point at which left-censoring occurs will be within one of several regions of analytical measurement. For example, Keith *et al.* (1983) defined three regions of measurement: the region of high uncertainty, the region of less-certain analysis, and the region of quantitation. The region of high uncertainty lies below the DL. The region of less-certain quantitation has a lower limit at the DL and an upper limit at the concentration at which the certainty level for the sample concentration is ±30 percent (confidence level=99 percent). This upper limit is called the limit of quantitation (LOQ). The region of quantitation is contained within the region of linearity of the calibration curve above the DL. Keith *et al.* (1983) suggested that "quantitative interpretation, decision-making and regulatory actions should be limited to data at or above the limit of quantitation." Thus, left-censoring could occur at the LOQ. Their recommendation that the LOQ be considered the data-reporting limit (Gilliom *et al.* 1984) is laudable when only the analytical process is considered. However, legal constraints on methodology, inclusion of control or "upstream" sampling sites, and wide temporal fluctuations in chemical variables often produce data sets with values in all regions of measurement. Their recommendation that data be eliminated from statistical interpretation because a subset of sites or samplings was characterized by values less than the LOQ would severely limit the utility of the data set (Porter *et al.* 1988). In contrast to Keith *et al.* (1983), Gilbert and Kinnison (1981), Gilliom *et al.* (1984), and Porter *et al.* (1988) advocate the retention of values less than the DL in data sets to avoid unnecessary loss of information. Other workers have defined less extreme methods of coping with this source of uncertainty. Most involve left-censoring of data sets at the DL.

The nature of left-censoring can vary within and between data sets. Types of censoring originally defined for life or failure testing (right-censoring) are also used to categorize types of left censoring. In Type I censoring, the point of censoring is fixed for all observations and the number of censored observations varies. The most common occurrence of left-censoring of environmental data sets involves samples that have a common

[1]Paper No. 88143 of the *Water Resources Bulletin*. Discussions are open until April 1, 1990.
[2]Respectively, Assistant Ecologist, University of Georgia, Savannah River Ecology Laboratory, P.O. Drawer E, Aiken, South Carolina 29801; Assistant Ecologist, Savannah River Ecology Laboratory; Research Engineer, E.I. DuPont de Nemours & Co., Savannah River Laboratory, Atomic Energy Division, Savannah River Plant, Aiken, South Carolina 29801; and Associate Ecologist, Savannah River Ecology Laboratory.

and well-defined point of censoring, the DL. In contrast, Type II censoring is characterized by random points of censoring and a fixed portion of the observations censored. Gilbert and Kinnison (1981) give the following example of Type II censoring for assays of environmental radioactivity. Type II censoring applies if, before an analytical session, counting times are set using prior understanding of the range of activities in the sample set such that a fixed proportion of all observations will remain undetected. This may be done for time management as some samples may require excessively long counting times for adequate quantification. Methods of coping with these two types of censoring will be defined in this report.

Unfortunately, more complex forms of Type I or II censoring often occur in data sets from environmental monitoring. These complex forms of censoring are designated herein as multiple detection limit censoring (MDLC); they are characterized by a randomly or monotonically varying DL between subsets of observations within the data set. For example, five sites on a stream were sampled monthly for a period of five years. Each time that samples were taken, a DL was calculated. However, analytical factors produced a significantly different DL for each sampling event. Therefore, when the entire sample set was analyzed statistically (n=300), each subset of five observations (5 sites/sampling) had a different DL from other subsets in the data set. Alternatively, the DL over the five years of sampling could have decreased with improved methods. In this case, the DL for each sampling subset would decrease in a monotonic fashion with increased duration of sampling. MDLC is similar to progressive censoring as defined for right censoring in this case. In our opinion, MDLC occurs often in environmental data sets; however, the results are often treated as singly censored data of Type I. Unfortunately, the problems associated with treating MDLC data as singly censored Type I or Type II, or progressively censored data are only presently being defined (Helsel and Cohn, 1988).

Statistical techniques to accurately handle censored data have been developed during the last thirty years but are not yet widely used for environmental data. Many analysts still omit values below the DL or replace them with the DL, DL/2 or 0 prior to calculating means and variances. The statistical properties of several estimators of means and variances have been recently reviewed by Schneider (1986) and Porter *et al.* (1988), but they did not present a decision-maker possessing an average statistics background with easily accessible methods of dealing with censored data sets. Gleit (1985), and Gilliom and Helsel (1986) have recently examined the properties of some techniques in a statistically rigorous manner that are not intended to provide the non-statistician with easily implemented procedures.

The present work identifies and illustrates the most generally useful of several current methods for dealing with both Type I and II censored data. Recommendations derived by simulations (Gleit, 1985; Gilliom and Helsel, 1986) will be discussed. The text in combination with a Pascal program (UNCENSOR) provides the non-statistician with the tools necessary to effectively estimate the mean and variance of data sets containing below DL values. Further, potential problems associated with MDLC and current guidelines for estimating associated means and variances will be discussed.

METHODS

General

Three water quality data sets were selected to represent typical situations: a normally distributed sample with many observations, a log-normally distributed sample with many observations and a normally distributed sample with few observations. All data were produced by analyses deemed "in control" by a rigorous quality control/quality assurance program. None of the observations in the three data sets were below the DL. The statistical consequences of increasingly poor DL were illustrated by artificially censoring each data set. Any value below an arbitrary DL was replaced with "below DL". Means and variances of each data set were estimated by the following methods:

1. using only values above the DL,
2. replacing values below DL with 0,
3. replacing values below DL with DL/2,
4. replacing values below DL with the DL,
5. regression on expected order statistics (ROS),
6. maximum likelihood estimation (MLE),
7. one-step restricted MLE,
8. bias-corrected MLE.

"Detection limits" were arbitrarily chosen to censor from 0% to 60% of the values in each data set. The relative errors of the estimates of mean and standard deviation were then compared over the range of censoring intensities. Relative error was defined as 100* [(Censored-Noncensored Estimate)/(Censored Estimate)].

Choice of Methods

At least two criteria can be used to choose a good estimator: bias and mean-squared error (MSE). Both can

be evaluated by simulating many random data sets and computing an estimate from each one. Bias is the difference between the true value and the average estimated value. Mean-squared error (Variance + Bias^2) measures overall accuracy by generalizing the variance. Numerous estimators have been proposed to estimate the mean or variance from censored, normally distributed data (see discussion). Many methods were developed as short-cut approximations to simplify hand calculations; other techniques were developed for Type II censoring. For this reason, they may not be appropriate for Type I censored data. Schneider (1986) described several of the methods for estimating parameters of censored, normal distributions. Four methods with low bias and MSE were selected based on his simulations of Type II censoring.

In general, maximum likelihood estimators (MLEs) have good statistical properties. MLEs for Type I censored normal distributions were derived by Cohen (1950); they do not have a closed-form solution and must be solved iteratively or by using a table (Cohen, 1959). They have the smallest mean-squared error of the available techniques, even for samples with small numbers of observations (Harter and Moore, 1966). The iterations required to calculate Cohen's estimator can be avoided by finding the solution to a restricted likelihood function, which can be solved directly (Persson and Rootzen, 1977). Both the iterative MLE and the restricted MLE are biased in small samples but the amount of the bias can be reduced by using a correction factor (Saw, 1961; Schneider, 1986). The use of this bias correction is made at the cost of a slightly higher MSE.

An alternative group of estimators uses expected values of normal order statistics to estimate the mean and variance of the data set. Conceptually, the simplest method involves replacement of censored observations with their predicted value from a linear regression of observed values on expected order scores (Gilliom and Helsel, 1986). Other permutations of this procedure estimate the mean and standard deviation by the intercept and slope of a regression of the observed data on expected order statistics (Barnett, 1975), and re-express the regression calculations as linear combinations of the observed values (Gupta, 1952; Sarhan and Greenberg, 1956). The order statistic methods are unbiased but have higher MSEs than do the maximum likelihood estimators. Many of the order statistics methods were designed to simplify calculations prior to the advent of widely available computer resources.

Software

Estimation methods described herein have been incorporated into a Pascal program, UNCENSOR. This program provides estimates for normal and two-parameter, log-normal data sets. Large ($n > 20$) and small ($n \leq 20$) data sets can be handled. UNCENSOR provides estimates of mean and variance by the application of several order statistics and maximum likelihood methods. At present, these methods include the following: iterative maximum likelihood method (Cohen, 1959), restricted, one-step, maximum likelihood method (Saw, 1959; Schneider, 1986), iterative order statistic method ("fill-in" with expected value, Gleit, 1985) and regression order statistic method (Looney and Newman, in prep.). A small sample size-bias corrected, maximum likelihood method (Saw, 1959; Schneider and Weissfeld, 1986) and log transformation bias correction (Aitchison and Brown, 1957) are available. These methods are described in more detail in Appendix A.

UNCENSOR requires an IBM PC or close compatible with at least 256K of RAM and a single floppy disk drive. It will check for and make use of an 8087 math coprocessor if it is available. It uses a Pull Down Window Interface (PDWI). Data may be entered as summary data (example: total number of samples, number below the DL, DL, mean of uncensored observations, variance of uncensored data) or raw data ($n \leq 1000$) from an ASCII file. Log transformation and sorting of data sets are performed by UNCENSOR when required. Output can be sent to the screen or a printer. Output options include a debug option in which intermediate values and results of calculations are provided. This public domain program is available from the senior author.

RESULTS

Table 1 lists the characteristics of the three data sets used to evaluate these estimation techniques. The null hypothesis of normality (or normality of log-transformed values) was not rejected at $\alpha = 0.10$ for these data sets.

Total alkalinity concentrations ($n = 135$) in the Savannah River adjacent to the Savannah River Plant (SRP) appeared to be normally distributed. The replacement or deletion estimators (#1-#4) gave biased estimates of the mean and standard deviation at low intensities of censoring. When the values below the DL were replaced with either 0 or DL/2, the estimates became increasingly biased downward as the intensity of censoring increased. This resulted from the larger difference between the mean of the observed values and the replacement value (0 or DL/2) (Figure 1). The estimates became increasingly biased upward when censored observations were deleted or replaced by the DL. Relative errors of the standard deviation estimates

TABLE 1. Summary of Data Sets Used for Illustrating the Effects of Data Censoring.

Variable	Site	N	Mean	Variance	W or D	P for D or W
Total alkalinity (mg/L as $CaCO_3$)	Savannah River	135	18.3	14.7	0.0709(D)	0.094
Total iron	Savannah River					
(µg/L)		117	997	396		
(ln µg/L)		117	6.67	0.25	0.0724(D)	0.016
Sulfate (mg SO_4/L)	Pen Branch Creek	20	5.21	4.65	0.9670(W)	0.990

N = sample size
W and D = the statistic for the test of random sampling from a normal distribution
P = the probability associated with the null hypothesis that the sample was taken from a normal population. Small values of W and large values of D imply rejection of the null hypothesis.

show the opposite pattern; replacing censored observations with 0 or DL/2 leads to estimated standard deviations that are too large, while deletion or replacement with DL underestimates the standard deviation. The use of 0 in place of the below DL values produced the most biased estimates of the mean and standard deviation.

The relative errors of the three maximum likelihood estimators were very similar and were small relative to those of other techniques (Figure 1). All three MLEs were essentially unbiased when the DL was less than 15 mg/L as $CaCO_3$ (22 percent censoring intensity), and the biases were relatively small as along as the DL was below the mean (less than 50 percent censoring intensity). If the DL was close to the mean, all three estimates of the mean are slightly high and all three estimates of the standard deviation are slightly low. The ROS estimator (#5) performed nearly as well as the MLEs. It was essentially unbiased when the DL was less than 15 mg/L as $CaCO_3$, but it was more biased than the MLEs as the censoring intensity increased.

Tests of normality of nontransformed and log-transformed values of total iron in the Savannah River (n = 117) suggested that these data could be best fitted by a 2-parameter, log-normal distribution. These data were log-transformed and then censored as described for the total alkalinity data set. After the mean and standard deviation of the log-transformed data were estimated, equations 8 and 9 were used to compute estimates of the mean and standard deviation of the untransformed data. The maximum likelihood estimators of the mean were the least biased, as noted for the alkalinity data, but no estimator of the standard deviation was best at all censoring levels (Figure 2). The MLEs were generally the least biased estimators of the standard deviation, except when a large fraction of the observations were below the DL.

Sulfate concentrations in Pen Branch Creek, a thermal stream on the SRP were used to test estimator performance in a small data set (n = 20, Figure 3). As with the total alkalinity data set, the ML and ROS estimators were less biased than the deletion or replacement estimators. In this small data set, the bias-corrected ML and ROS estimates of the mean were less biased, but the estimated standard deviations were more biased than the other two MLEs. Both the ROS and bias-corrected ML estimators of standard deviation had the undesirable property of being biased in noncensored samples.

DISCUSSION

Quantitative interpretation of data sets is most effective when all values are greater than the LOQ. Therefore, an ultimate goal in analytical method selection and experimental design selection should be the generation of a data set with all values in the region of quantitation. When it is not possible to generate a data set with all values above the LOQ (or the DL), the "no censoring rule" (Gilbert and Kinnison, 1981; Gilliom *et al.*, 1984; Porter *et al.*, 1988) is recommended during initial data archiving. The distribution of values within the various regions of measurement should be presented clearly when reporting results if values are obtained from more than one region of measurement. As discussed by Porter *et al.* (1988), estimation should be made of system error in conjunction with reporting data under the "no censoring rule".

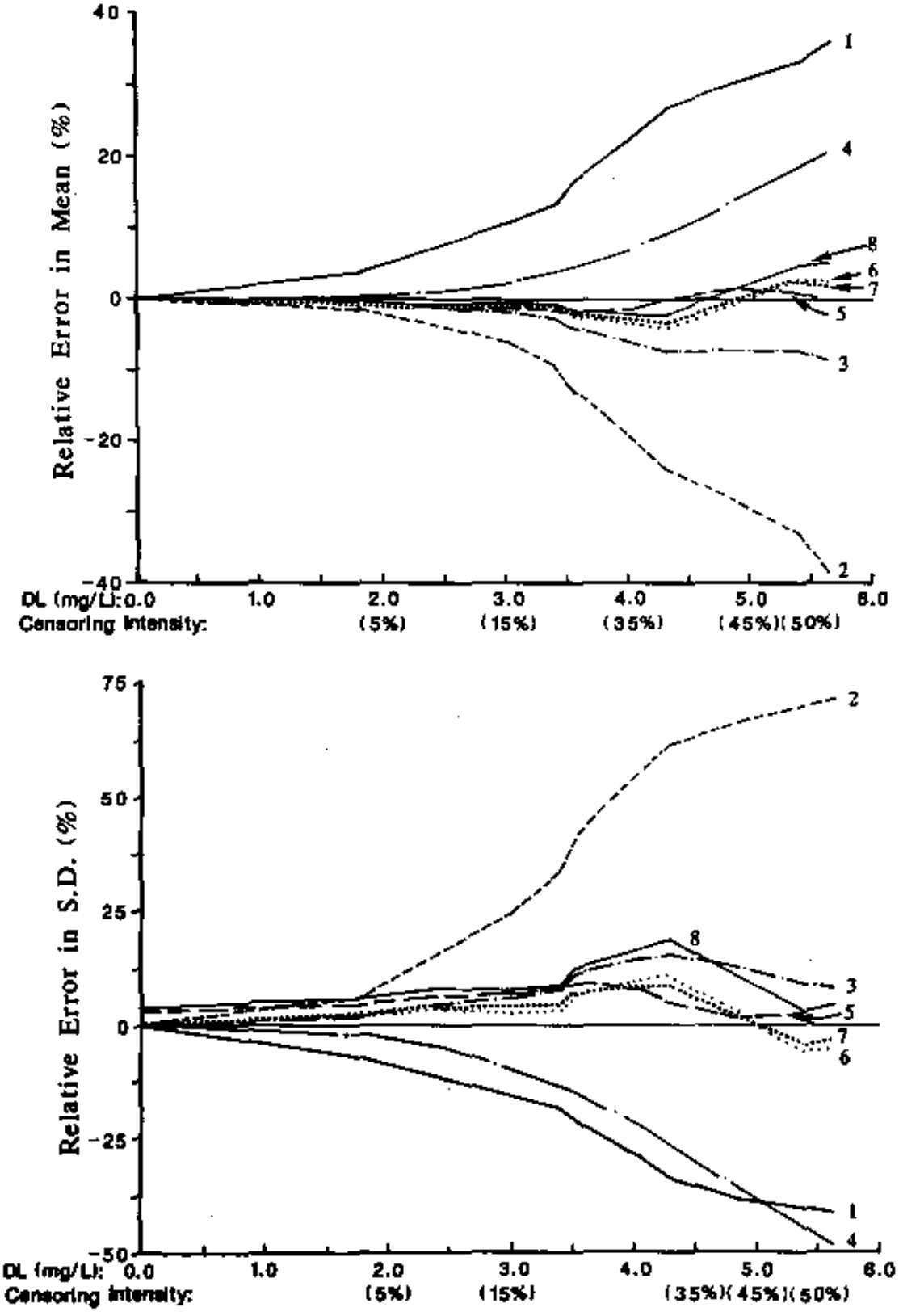

Figure 1. The Comparative Behavior of Deletion, Substitution and Statistical Methods of Estimating Mean and Standard Deviation of a Normally Distributed, Total Alkalinity Data Set (n = 135) That Was Left-Censored with Increasing Intensity.
Deletion: 1 ————; Enter as 0: 2 – – – – –; Enter as DL/2: 3 — - — - —; Enter as DL: 4 —— - ——; Regression on Expected Order Statistics: 5 — — — —; Iterative MLE: 6 ——————; 1-Step Restricted MLE: 7 – – – – – –; and Bias-Corrected MLE: 8 ————;

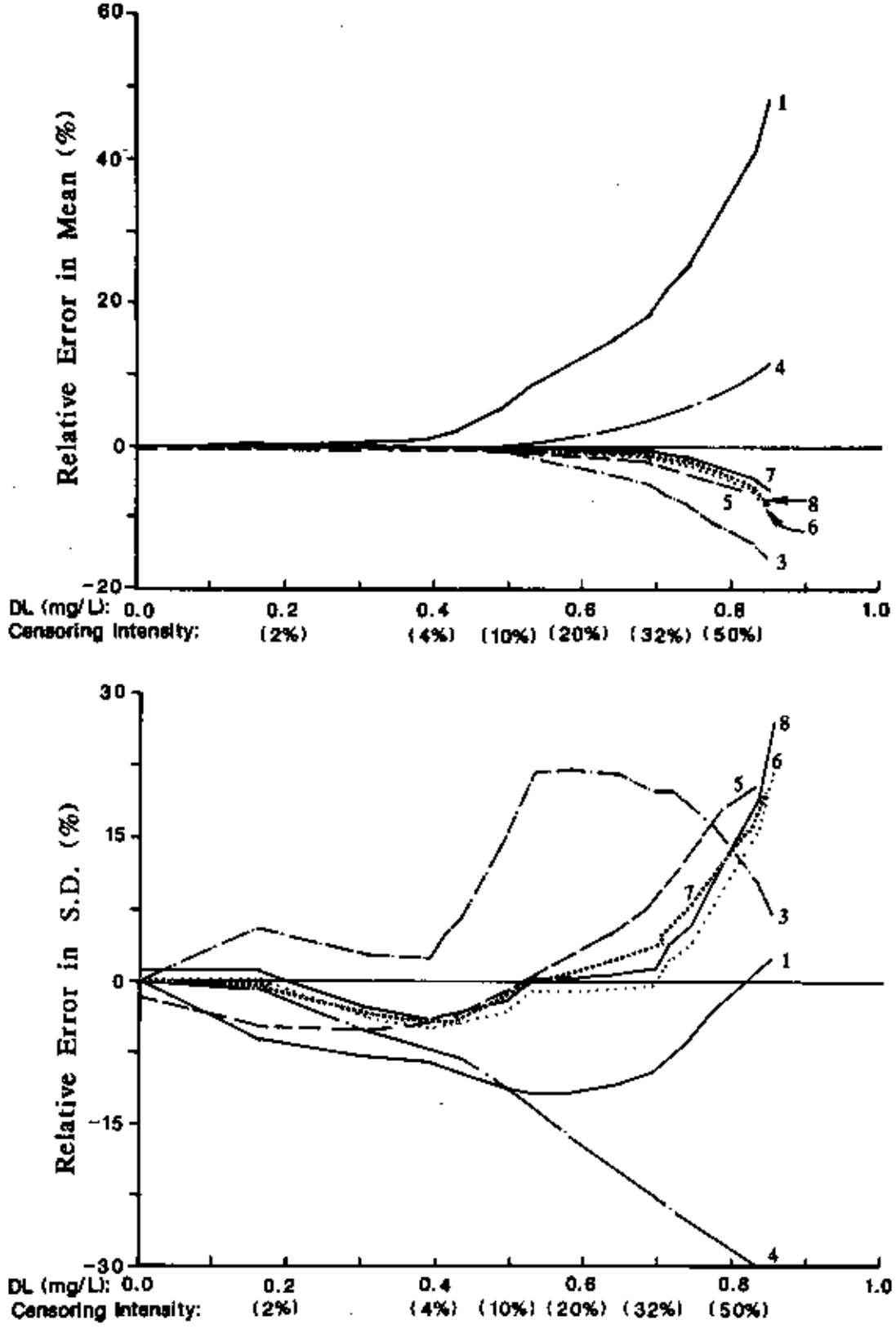

Figure 2. The Comparative Behavior of Deletion, Substitution, and Statistical Methods of Estimating Mean and Standard Deviation of a Log-Normally Distributed, Total Iron Data Set (n = 117) That Was Left-Censored with Increasing Intensity.
Deletion: 1 ————; Enter as 0: 2 – – – – –; Enter as DL/2: 3 — · — · —; Enter as DL: 4 —— · ——; Regression on Expected Order Statistics: 5 — — — —; Iterative MLE: 6 ————; 1-Step Restricted MLE: 7 – – – – – –; and Bias-Corrected MLE: 8 ————;

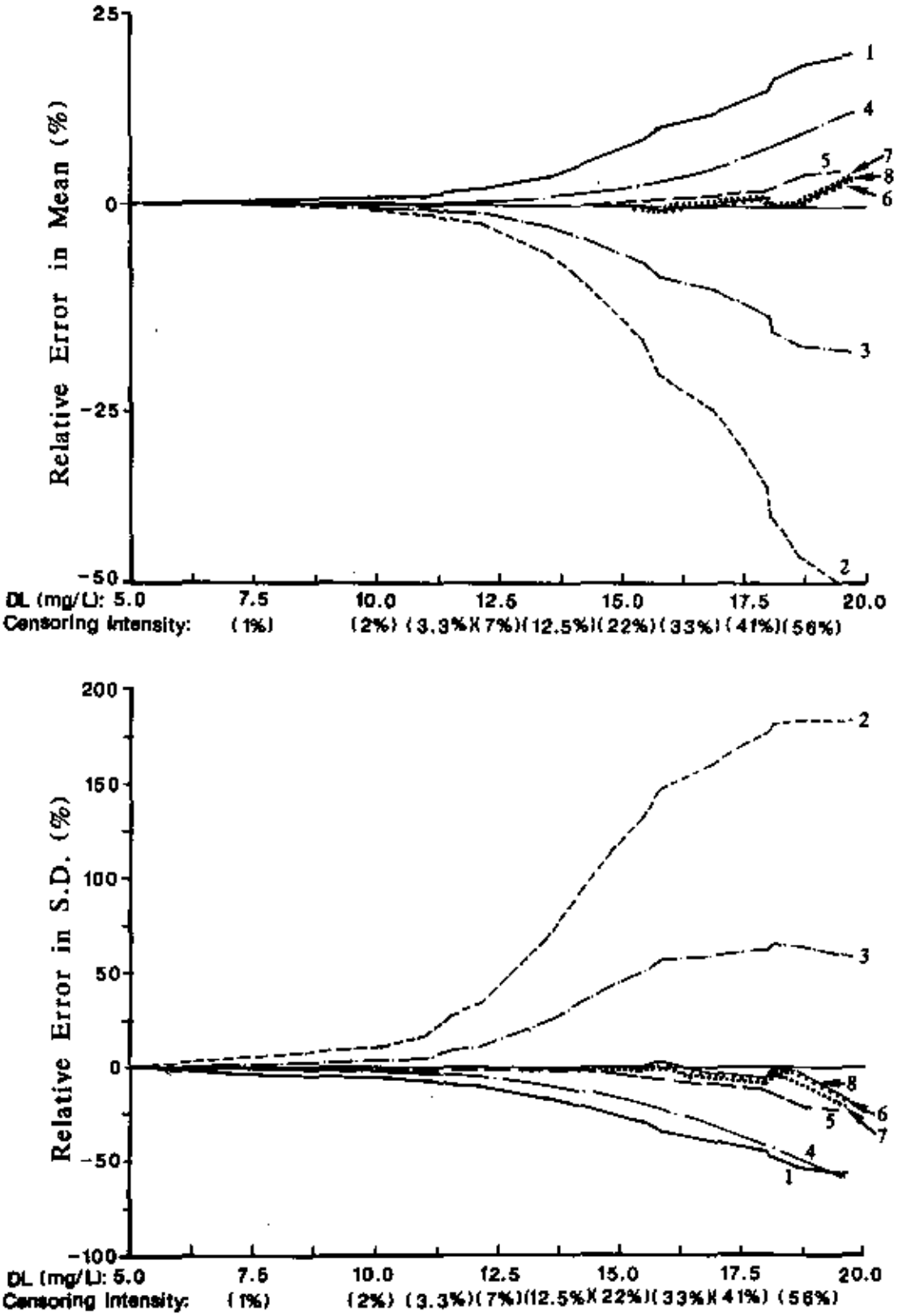

Figure 3. The Comparative Behavior of Deletion, Substitution, and Statistical Methods of Estimating Mean and Standard Deviation of a Normally Distributed, Sulfate Data Set (n = 20) That Was Left-Censored with Increasing Intensity.
Deletion: 1 ————; Enter as 0: 2 – – – – – –; Enter as DL/2: 3 — · — · —; Enter as DL: 4 —— · ——; Regression on Expected Order Statistics: 5 — — — —; Iterative MLE: 6 ·····; 1-Step Restricted MLE: 7 – – – – – –; and Bias-Corrected MLE: 8 ————;

When only the censored data set is available, omission of values below the DL or use of 0, DL/2, or DL in place of values below the DL produces biased estimates of mean and variance during subsequent statistical analysis. The intensity of the bias will worsen as the degree of censoring increases. Therefore, the use of these techniques is not recommended.

When knowledge of the underlying distribution for the censored data set is available, the mean and variance can be estimated for censored data sets. Clearly, the value of the estimates of mean and variance of censored or noncensored data sets will depend on the validity of the assumptions regarding the underlying distribution. As the intensity of censoring increases, the information available for establishing the nature of the underlying distribution decreases.

Gilliom and Helsel (1986), and Helsel and Gilliom (1986) correctly question the ease of assigning a censored water quality data set to a specific parent distribution. These authors examined the behaviors of several estimation methods for censored data with unknown parent distributions or for censored data sets with misclassified distributions. The most robust estimation method for estimating the mean and variance was least squares regression of normal scores after log transformation of the observations (ROS or Gleit, 1985). MLE methods performed poorly at high intensities of censoring. Difficulties associated with accurate classification are dependent on the sample size, censoring intensity, and relative quartile range (rqr) (Gilliom and Helsel, 1986). When insufficient information is available regarding the parent distribution, these robust estimation methods should be employed.

Numerous statistical methods are available for estimation of mean and variance of Type I- and Type II-censored data sets (Table 1). Maximum likelihood and order statistics methods for normal and 2-parameter, log-normal distributions are presented here. Further discussion of these and other methods for normally distributed, censored data sets can be found in the works of Stevens (1937), Hald (1949), Cohen (1950, 1957, 1959, 1961), Gupta (1952), Sarhan and Greenberg (1956, 1958, 1962) Saw (1959, 1961) and Schneider (1986). Detailed discussions of estimation methodologies for 2-parameter, log-normally distributed data sets are available in Aitchison and Brown (1957), Kushner (1976), and Gilbert and Kinnison (1981). Similar discussions of 3-parameter, log-normally distributed data sets are found in Spiller (1948), Harter and Moore (1966), Munro and Wixley (1970), and Gilbert and Kinnison (1981). Watterson (1959) extended the methods of Gupta (1952), and Sarhan and Greenberg (1956, 1958) to include censored, multivariate samples. Cohen *et al.* (1978) provided maximum likelihood methods for left-censored Weibull and Gamma distributions.

Selection of any of the above methods depends on the underlying distribution, number of samples, nature of censoring, acceptable levels of bias and efficiency, intensity of censoring and computational ease. For those writing their own computer programs or performing calculation by hand, we suggest the use of the restricted MLE (Estimator #7) (Persson and Rootzen, 1977) as an easily computed and statistically well behaved estimator of mean and variance of normal and 2-parameter, log-normal data sets when the underlying distribution is known. These methods are not recommended for mean and variance estimation when data classification to a parent distribution is ambiguous.

Methods for estimating mean and variance for multiple detection limit censoring of environmental data sets are still being established (Helsel and Cohn, 1988). Acceptability of direct application of methods for singly censored data was assessed and rejected as compromised during Monte Carlo simulations. Considerable information was lost when the highest DL was used as the single threshold of censoring in these simulations. Helsel and Cohn (1988) evaluated probability plotting (Hirsch and Stedinger, 1986), maximum likelihood (Cohen, 1976) and adjusted maximum likelihood (Cohen, 1988) methods of estimating mean and variance for MDLC log-normal data sets. The adjusted maximum likelihood method had the lowest root-mean-square error (rmse); but, when misspecification of the parent distribution was likely, the robustness of the probability plotting method was superior to the other methods.

Other methods developed for multiple censored life testing data have not been assessed completely for use with MDLC data. Type I progressively censored data sets arise when subsets of items are removed from the life test in a series of time intervals (T_i). Each subset has an increasing maximum time to failure as the experiment progresses. Time intervals are fixed, and the number of survivors is random (Cohen, 1975). For left-censoring as discussed herein, Type I progressive censoring would involve a fixed DL for each subset and a random number of observations below the DLs. Type II, right-censoring of similar data, has been defined in the statistical literature as multiple censored (Herd, 1956) or hypercensored (Roberts, 1962) samples. In this type of censoring, the number of survivors is fixed, and the time intervals (T_i) are random (Cohen, 1975). For left-censoring as discussed herein, Type II progressive censoring would have random DL with the number of observations below each DL fixed. If the parent distribution is known, a rich literature is available for consideration. For example, Cohen discusses maximum likelihood methods for progressively censored life testing data sets from normal (Cohen, 1963), exponential (Cohen, 1963), 3-parameter Weibull (Cohen, 1975), 3-parameter log-normal (Cohen, 1976) and 3-parameter

gamma (Cohen and Norgaard, 1977) distributions. Further assessment of these techniques for use with MDLC water quality data sets is needed. These methods would be especially valuable during statistical analysis of long-term monitoring data.

SUMMARY

1. An ultimate goal of analytical method development and experimental design should be the generation of data sets with all values above the LOQ.

2. The "no censoring rule" should be observed during data archiving when the data set contains values in several regions of measurement.

3. The omission of values below the DL or the use of 0, DL/2, or the DL in place of the values below the DL may produce biased estimates of mean and variance.

4. Several statistical methods are available for estimating means and variances for Type I- and II-censored data sets. Restricted maximum likelihood estimates are easily computable, are less biased and more accurate than the other estimators when the parent distribution is known. The robustness of regression of normal scores after log transformation of the data suggests that these methods are the most effective when a parent distribution cannot be identified.

5. UNCENSOR, a public domain Pascal program, is available for application of maximum likelihood and order statistics methods to normal and 2-parameter log-normal data sets.

6. Methods for coping with MDLC of environmental data sets remain ill-defined. Further work in this area is needed.

ACKNOWLEDGMENTS

This research was supported by contract DE-AC09-76SROO-819 between the U.S. Department of Energy and the University of Georgia. Jean B. Coleman drafted the figures, and Marianne Reneau typed the manuscript. Charles Segal and Dawn Givens developed the code for the program, UNCENSOR. The authors are grateful to Drs. A.C. Cohen, R.O. Gilbert, A. Gleit, and two anonymous reviewers who provided helpful comments. Drs. A.C. Cohen, A. Gleit, R.O. Gilbert, and M.E. Mulvey provided excellent reviews of early drafts of this manuscript. Some of the information in this manuscript was developed for the 1986 American Water Works Association Water Quality Technology Conference (November 16-20, 1986, Portland, Oregon).

LITERATURE CITED

Aitchison, J., and J.A.C. Brown, 1957. The Lognormal Distribution with Special Reference to Its Use in Economics. Cambridge University Press, New York, 176pp.

Barnett, V., 1975. Probability Plotting Methods and Order Statistics. Applied Statistics 24: 95-108.

Blom, G., 1958. Statistical Estimates and Transformed Beta Variables. John Wiley and Sons, Inc., New York, 176 pp.

Cohen, A.C., Jr., 1950. Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples. Annals of Mathematical Statistics 21: 557-569.

Cohen, A.C., Jr., 1957. On the Solution of Estimating Equations for Truncated and Censored Samples from Normal Populations. Biometrika 44: 225-236.

Cohen, A.C., Jr., 1959. Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated. Technometrics 1: 217-237.

Cohen, A.C., Jr., 1961. Tables of Maximum Likelihood Estimates: Singly Truncated and Singly Censored Samples. Technometrics 3: 535-541.

Cohen, A.C., Jr., 1963. Progressively Censored Samples in Life Testing. Technometrics 5: 327-339.

Cohen, A.C., Jr., 1975. Multi-Censored Sampling in the Three Parameter Weibull Distribution. Technometrics 17: 374-351.

Cohen, A.C., Jr., 1976. Progressively Censored Sampling in the Three Parameter Log-Normal Distribution. Technometrics 18: 99-103.

Cohen, A.C., Jr. and N.J. Norgaard, 1977. Progressively Censored Sampling in the Three-Parameter Gamma Distribution. Technometrics 19: 333-340.

Cohen, A.C., Jr., K.S. Crump, W.B. Smith, and C.M. Dayton, 1978. Statistical Analysis of Radionuclide Levels in Food Commodities. Report to the U.S. Department of Health, Education and Welfare, Food and Drug Administration, Division of Mathematics.

Gilbert, R.O. and R.R. Kinnison, 1981. Statistical Methods for Estimating the Mean and Variance from Radionuclide Data Sets Containing Negative, Unreported or Less-Than Values. Health Physics 40: 377-390.

Gilliom, R.J. and D.R. Helsel, 1986. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 1. Estimation Techniques. Water Resource Research 22: 135-146.

Gilliom, R.J., R.M. Hirsch, and E.J. Gilroy, 1984. Effect of Censoring Trace-Level Water-Quality Data on Trend-Detection Capability. Environmental Science and Technology 18: 530-539.

Gleit, A., 1985. Estimation for Small Normal Data Sets with Detection Limits. Environmental Science and Technology 19: 1201-1206.

Gupta, A.K., 1952. Estimation of the Mean and Standard Deviation of a Normal Population from a Censored Sample. Biometrika 39: 260-273.

Hald, A., 1949. Maximum Likelihood Estimation of the Parameters of a Normal Distribution Which Is Truncated at a Known Point. Skandinavisk Aktuarietidskrift 32: 119-134.

Harter, H.L. and A.H. Moore, 1966. Local-Maximum-Likelihood Estimation of the Parameters of Three-Parameter Lognormal Populations from Complete and Censored Samples. Journal American Statistical Association 61: 842-851.

Helsel, D.R. and T.A. Cohn, 1988. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resource Research 24: 1997-2004.

Helsel, D.R. and R.J. Gilliom, 1986. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 2. Verification and Applications. Water Resource Research 22: 147-155.

Herd, G.R., 1956. Estimation of the Parameters of a Population from a Multi-Censored Sample. Ph.D. Dissertation, Iowa State College.

Hirsch, R.M. and J.R. Stedinger, 1987. Plotting Positions for Historical Floods and Their Precision. Water Resource Research 23: 715-727.

Johnson, N.L. and S. Kotz, 1969. Distributions in Statistics (Vol. 1). Discrete Distributions, Houghton Mifflin, Boston, 323 pp.

Keith, L.H., W. Crummett, J. Deegan, R. Libby, J.K. Taylor and G. Wentler, 1983. Principles of Environmental Analysis. Analytical Chemistry 55: 2210-2218.

Kushner, E.J., 1976. On Determining the Statistical Parameters for Pollution Concentration from a Truncated Data Set. Atmospheric Environment 10: 975-979.

Mandel, J., 1964. The Statistical Analysis of Experimental Data. Interscience Publishers, New York, 410 pp.

Munro, A.H. and R.A.J. Wixley, 1970. Estimators Based on Order Statistics of Small Samples from a Three-Parameter Lognormal Distribution. Journal American Statistics Association 65: 212-225.

Persson, T. and H. Rootzen, 1977. Simple and Highly Efficient Estimators for a Type I Censored Normal Sample. Biometrika 64: 123-128.

Porter, P.S., R.C. Ward and H.F. Bell, 1988. The Detection Limit. Water Quality Monitoring Data Are Plagued with Levels of Chemicals That Are Too Low to Be Measured Precisely. Environmental Science and Technology 22: 856-861.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, 1986. Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge, 818 pp.

Roberts, H.R., 1962. Some Results in Life Testing Based on Hypercensored Samples from an Exponential Distribution. Ph.D. Dissertation, George Washington University.

Sarhan, A.E. and B.G. Greenberg, 1956. Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples. Part I. The Normal Distribution Up to Samples of Size 10. Annals of Mathematical Statistics 29: 79-105.

Sarhan, A.E. and B.G. Greenberg, 1958. Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples. Part II. Tables for the Normal Distribution for Samples of Size $11 \leq n \leq 15$. Annals of Mathematical Statistics 29: 79-105.

Sarhan, A.E. and B.G. Greenberg (eds.), 1962. Contributions to Order Statistics, John Wiley and Sons, Inc., New York, 482 pp.

Saw, J.G., 1959. Estimation of the Normal Population Parameters Given a Singly Censored Sample. Biometrika 46: 150-159.

Saw, J.G., 1961. The Bias of the Maximum Likelihood Estimates of Location and Scale Parameters Given a Type II Censored Normal Sample. Biometrika 48: 448-451.

Schneider, H., 1984. Simple and Highly Efficient Estimators for Censored Normal Samples, Biometrika 71: 412-414.

Schneider, H., 1986. Truncated and Censored Samples from Normal Populations. Marcel Dekker, New York, 273 pp.

Schneider H. and L. Weissfeld, 1986. Inference Based on Type II Censored Samples. Biometrics 42: 531-536.

Spiller, D., 1948. Truncated Log-Normal and Root-Normal Frequency Distributions of Insect Populations. Nature 162: 530-531.

Stevens, W.L., 1937. The Truncated Normal Distribution, Annals of Applied Biology 24: 847-852.

Watterson, G.A., 1959. Linear Estimation in Censored Samles from Multivariate Normal Populations, Annals of Mathematical Statistics 30: 814-824.

APPENDIX A: METHODS USED HEREIN

We selected four estimators that had low bias and low MSE in Schneider's simulation (Schneider, 1986). The four estimators included three MLE methods and one order statistics method. As discussed below, MLE methods are based on the statistical properties of the noncensored portion of the data, adjusted for the theoretical effects caused by the defined intensity of censoring. Order statistics are based on the behavior of the noncensored portion of the data assuming an underlying normal probability density function for the entire data set. All of these methods explicitly rely on the acceptability of the assumed probability distribution.

A variety of order statistics methods are available for data sets censored at the DL. While all of these methods are fundamentally similar, we demonstrate a regression on expected order statistics (ROS) method because it is conceptually straightforward, general in nature, and essentially the same estimator as recommended by Gleit (1985) in his simulation study. Observations are ranked from smallest to largest with values below the DL treated as the smallest values. Let

n = total number of observations,
k = number of observations below the DL,
X_i = the i'th ranked observation,
X_1 = the smallest value,
X_{k+1} = the smallest detectable value, and
x_n = the largest value.

If the observations comprising the sample are randomly drawn from the population, the ordered data values would divide the underlying probability density function into equal areas. Thus, an estimated plotting position on an appropriate coordinate system can be calculated for each point such that the noncensored portion of the data will fall on a straight line. Specifically, we calculate $A_i = F^{-1}[(i - 3/8)/n + 1/4)]$ where $F^{-1}[x]$ is the inverse cumulative normal distribution function (Blom, 1958; Mandel, 1964; Press *et al.* (1986). Consider only the n-k points above the DL and regress X_i on A_i to estimate a and b in the equation,

$$X_i = a + bA_i + e_i . \tag{1}$$

The mean of the noncensored distribution is estimated by a and the standard deviation is estimated by b.

The maximum likelihood estimators of a Type I censored normal distribution were derived by Cohen (1950, 1959). The MLEs of the mean and standard deviation are those values that solve the system of equations:

$$\hat{\mu} = \bar{x} - \hat{\sigma}\left(\frac{k}{n-k}\right)\left(\frac{f(\varepsilon)}{F(\varepsilon)}\right) \tag{2}$$

$$\hat{\sigma}^2 = [S^2 + (\bar{x} - \mu^2)] \Big/ \left[1 + \varepsilon\left(\frac{k}{n-k}\right)\left(\frac{f(\varepsilon)}{F(\varepsilon)}\right)\right] \tag{3}$$

Where:

$$\varepsilon = \frac{DL - \mu}{\sigma} \quad (4)$$

f(x) = the distribution function for the normal distribution,
F(x) = the cumulative distribution function for the normal distribution,
$\bar{x}$ = the mean of all values above the DL, and
S = the population standard deviation of all values above the DL.

This system has no closed-form solution and must be solved iteratively. Choose suitable starting values for μ and σ such as $\bar{x}$ and S, compute ε using (4), and use (2) and (3) to compute new estimates of μ and σ, respectively. Iterate this process until the change in the estimates is less than some predetermined criterion.

Explicit solutions to the MLE equations can be obtained by imposing a small restriction (Persson and Rootzen, 1977). The number of observations with values below the DL has a binomial distribution with parameters, n and F(ε) (see definitions above). Using the properties of the binomial distribution, a natural estimate of F(ε) is k/n, hence ε can be estimated by:

$$\hat{\varepsilon} = F^{-1}\left(\frac{k}{n}\right) \quad (5)$$

The restriction to the MLEs is the replacement of ε in equations 2 and 3 with $\hat{\varepsilon}$. These restricted equations can be solved explicitly to give the following (Persson and Rootzen, 1977):

$$\hat{\mu}_{RML} = \bar{x} - a\sigma^* \quad (6)$$

$$\hat{\sigma}_{RML} = S^2 - (a\hat{\varepsilon} - a^2)(\sigma^*)^2 \quad (7)$$

where:

$a = F(\hat{\varepsilon})(n/k)$,
$C = \hat{\varepsilon}(\bar{x} - DL)$,
$\sigma^* = \frac{1}{2}[C + \sqrt{C^2 + 4S^2 + 4(\bar{x} - DL)^2}]$,
F(x) = normal probability density function, and
$\hat{\varepsilon}$ = defined in (5) above.

Both the iterative MLEs and the restricted MLEs are biased, but they have lower mean-squared errors than other, unbiased estimators. The low mean-squared error makes these estimators attractive to a statistician, but many users prefer unbiased estimators. The bias in the MLEs can be reduced by finding an equation to approximate the bias and using it to compute a corrected estimate. Bias approximations for Type II censored normal samples have been derived in Saw (1961), Schneider (1986), and Schneider and Weissfeld (1986). The bias corrected MLEs are given by the following:

$$\hat{\mu}_{BC} = \hat{\mu} - \frac{\hat{\sigma}}{n+1} B(\hat{\mu}, n, k) \quad (8)$$

$$\hat{\sigma}_{BC} = \hat{\sigma} + \frac{\hat{\sigma}}{n+1} B(\hat{\sigma}, n, k) \quad (9)$$

where the bias in the mean (B ($\hat{\mu}$,n,k)) is approximately

$$e^{\left[2.692 - 5.439\left(\frac{n-k}{n+1}\right)\right]}$$

and B ($\hat{\sigma}$,n,k) is approximately

$$\left[0.312 + 0.859\left(\frac{n-k}{n+1}\right)\right]^{-2}$$

The expressions for the approximate bias were obtained by Schneider (1986) by least-squares fitting of Saw's tables (1961). Although derived for Type II censoring, these bias corrections should also reduce the bias of Type I censored samples.

Mean and variance estimators for Type I-censored, 2-parameter log-normally distributed variables were obtained by using the above methods on log-transformed data. The resulting estimates were used in the following equations as described in Gilbert and Kinnison (1981):

$$\hat{\mu} = e^{\hat{\mu}_l}\,\psi_n(\hat{\sigma}_l^2/2) \quad (10)$$

$$\hat{\sigma}^2 = e^{2\hat{\mu}_l}\left[\psi_n(2\hat{\sigma}_l^2) - \psi_n\left(\frac{n-2}{n-1}\hat{\sigma}_l^2\right)\right] \quad (11)$$

where:

n = the number of observations,
$\hat{\mu}_l$ = the estimate of the mean for the log-transformed data,
$\hat{\sigma}_l$ = the estimate of the standard deviation for the log-transformed data, and
$\psi_n(t)$ = a value obtained from Table A2 of Aitchison and Brown (1957).

It is important to realize that a simple, back-transformation of mean and variance estimates from censored, log-transformed data will produce biased estimators of the arithmetic mean and variance of log-normally distributed data. The reader is referred to

Chapter 5 of Aitchison and Brown (1957) for further discussion of this point.