WATER RESOURCES RESEARCH, VOL. 22, NO. 2, PAGES 135–146, FEBRUARY 1986

Estimation of Distributional Parameters for Censored Trace Level Water Quality Data
1. Estimation Techniques

ROBERT J. GILLIOM AND DENNIS R. HELSEL

U.S. Geological Survey, Reston, Virginia

A recurring difficulty encountered in investigations of many metals and organic contaminants in ambient waters is that a substantial portion of water sample concentrations are below limits of detection established by analytical laboratories. Several methods were evaluated for estimating distributional parameters for such censored data sets using only uncensored observations. Their reliabilities were evaluated by a Monte Carlo experiment in which small samples were generated from a wide range of parent distributions and censored at varying levels. Eight methods were used to estimate the mean, standard deviation, median, and interquartile range. Criteria were developed, based on the distribution of uncensored observations, for determining the best performing parameter estimation method for any particular data set. The most robust method for minimizing error in censored-sample estimates of the four distributional parameters over all simulation conditions was the log-probability regression method. With this method, censored observations are assumed to follow the zero-to-censoring level portion of a lognormal distribution obtained by a least squares regression between logarithms of uncensored concentration observations and their *z* scores. When method performance was separately evaluated for each distributional parameter over all simulation conditions, the log-probability regression method still had the smallest errors for the mean and standard deviation, but the lognormal maximum likelihood method had the smallest errors for the median and interquartile range. When data sets were classified prior to parameter estimation into groups reflecting their probable parent distributions, the ranking of estimation methods was similar, but the accuracy of error estimates was markedly improved over those without classification.

INTRODUCTION

Interest in the occurrence of trace levels of toxic substances in surface and ground waters and their effects on human health and aquatic ecosystems has increased during the last 10 years. However, investigations of trace substances in ambient waters have encountered a recurring difficulty: a substantial portion of water sample concentrations are below the limits of detection established by analytical laboratories. Measurements below the detection limit are generally reported as "less than the detection limit" rather than as numerical values. Data sets with "less-than" observations are termed "censored data" in statistical terminology. Censored data do not present a serious interpretation problem if concentrations of primary interest are well above the detection limit, but this is often not the case. For some chemicals, established water quality criteria are below commonly applied detection limits. For many others, the great uncertainty in the effects of long-term exposure to very low levels also make it desirable to assess the frequency of occurrence of concentrations below the detection limit. In short, there is a need to estimate the frequency distribution of concentrations above, near, and below detection limits using only data above the detection limit.

The purpose of this study is to address several key aspects of estimating distributional parameters from censored data. These include (1) the performance of several estimation methods when estimating distributional parameters from small samples drawn from a wide range of underlying distributions and censored to varying degrees; (2) Criteria for determining, based only on attributes of data remaining after censoring, which estimation method is most likely to be best for each data set; and (3) the reliability of estimates from censored data of four distributional parameters: the mean, standard deviation, median, and interquartile range.

Paper number 5W4030.

PREVIOUS STUDIES

There have been extensive investigations of methods for estimating location and scale parameters for censored data drawn from specific parent distributions [*David*, 1981]. There have been far fewer studies of the application of these methods to environmental data for which parent distributions are unknown and sample sizes are small.

One of the first applications of censored data analysis in the environmental field was by *Leese* [1973], who applied censored data techniques to flood frequency analysis. She found that standard errors of mean annual flood estimates could be reduced by using the maximum likelihood estimates (MLE) for censored Gumbel distributions. Recently, *Condie and Lee* [1982] showed that maximum likelihood estimators for small censored samples from the three parameter lognormal and the log-Pearson type III distributions improved flood frequency estimates.

Owen and DeRouen [1981] addressed the problem of estimating a mean from censored air contaminant data. They used Monte Carlo techniques to evaluate the performance of MLE methods derived for lognormal and delta (lognormal augmented by some percentage of zeros) distributions when estimating the mean of censored data drawn from a combination of lognormal and delta distributions. For the range of sample sizes ($n = 5$ to $n = 50$), population coefficients of variation (CV $\cong$ 0.8–1.6), and degrees of type II censoring (5–25%) that they investigated, the delta MLE usually had lower mean square errors than the lognormal MLE. Type II censoring fixes the proportion of data censored in each data set, while type I censors all data below a fixed value [*David*, 1981]. Most recently, *Hashimoto and Trussell* [1983] compared several estimators of the mean for censored water quality data. Their examples illustrate the bias caused by three commonly used methods: discarding censored observations, setting all censored observations equal to zero, or assigning the detection limit to all censored observations. Their examples also included a comparison of estimates of the mean from the lognormal MLE to estimates made by filling in censored observations from a least squares regression relationship fit to uncensored observations plotted on a log-probability scale. That comparison suggested that the regression approach yields results very similar to those of the MLE method.

APPROACH

Generation of data. Sixteen parent distributions were selected as representative of the range of frequency distributions that is typical of trace water-quality data. Five hundred data sets of sample sizes 10, 25, and 50 observations were generated from each distribution. Each data set was censored at the 20th, 40th, 60th, and 80th percentiles of the parent distribution. Parameter estimation methods could then be evaluated for different sample sizes and degrees of censoring.

Parameter estimation methods. Eight methods were evaluated for estimating the mean, standard deviation, median, and interquartile range of censored data. The reliability and relative performance of methods was evaluated based on their root-mean-squared errors (rmses).

Estimation without classification. For each censoring level and sample size, all data sets from the 16 parent distributions were combined for computation of rmses for each method and distribution parameter. Best methods, based on minimum rmse, were identified for each parameter for every combination of censoring level and sample size. Rmses of these best methods for each such combination were evaluated in relation to the most robust method over all simulation conditions.

Estimation with classification. A goal was to improve method selection and the accuracy of rmses by classifying data sets based on attributes of data above the detection limit. Several sample statistics were computed for each data set and the one which best indicated the parent distribution was selected. Discriminant analysis by this variable determined criteria for identifying the most probable parent distribution(s) of a censored data set. All data sets were then classified using these criteria. Benefits in method selection and improved accuracies of rmses were evaluated.

GENERATION OF DATA

In designing the Monte Carlo experiments, a primary goal was to mimic as closely as possible the types of data that actually occur for concentrations of trace constituents in water. Hundreds of uncensored data sets for trace constituents were evaluated, including visual inspection of shapes and evaluation of the frequency distributions for the sample coefficients of variation (CV) and skewness. Coefficients of variation for 482 uncensored data sets (no measured concentrations were below the detection limit) for trace elements at U.S. Geological Survey river quality monitoring stations ranged from 0.15 to 3.2, with a median of 0.52. For the same data sets, sample skews ranged from −0.8 to 5.2 (6% were negative) with a median of 1.8.

Based on the sample properties and the visual inspection of sample histograms, four parent distributions with positive skew were chosen: lognormal, contaminated lognormal (mixture of two lognormals), gamma, and delta (lognormal augmented by zeros). Four variants of each distribution were considered, having CV's of 0.25, 0.50, 1.0, and 2.0. The resulting 16 parent distributions are herein abbreviated as LN(0.25), LN(0.50), LN(1.0), LN(2.0), CT(0.25), $\cdots$, GM(0.25), $\cdots$, DT(0.25), $\cdots$, DT(2.0). In all cases, the means equaled 1.0. The density function for each distribution is shown in Figure 1. The relationships used to generate data from these distributions are summarized below, followed by a brief description of the sizes and censoring of data sets. All x's refer to real space values and all y's refer to log space values.

Lognormal Distribution

When $y = \ln x$ is normally distributed with mean μ_y, and variance σ_y^2, a set of concentrations, x_i, $i = 1, \cdots, n$ can be generated using (1):

$$x_i = \exp(\mu_y + \sigma_y \varepsilon_i) \tag{1}$$

where ε_i is a randomly chosen value from a normal distribution with a mean of zero and variance of one.

Contaminated Lognormal Distribution

The contaminated lognormal distribution used in this study consists of a mixture of one predominant lognormal (μ_{x1}, σ_{x1}), which describes 80% of the overall population, and a contaminant lognormal (μ_{x2}, σ_{x2}), which describes 20% of the overall population. The approach to determining the characteristics of the two subpopulations was to specify proportional relationships between the parameters of the two distributions, which would allow unique solutions for their exact parameters for any overall distribution specified by μ_x and σ_x. The conditions imposed were $\mu_{x2} = 1.5\ \mu_{x1}$ and $\sigma_{x2}/\mu_{x2} = 2.0\ \sigma_{x1}/\mu_{x1}$. Under these conditions the relationships for μ_x and σ_x are given in the appendix.

Gamma Distribution

Two-parameter gamma distributions, characterized by a shape parameter, α_x, and a scale parameter, β_x were generated using the International Mathematical and Statistical Libraries generating routine.

Delta Distribution

The delta distribution is a mixture of a lognormal distribution (μ_{x1}, σ_{x1}) and some portion (p) of zero values. For all our simulations, the portion of zeros was 5% ($p = 0.05$). The mean and standard deviation of the overall distribution were given by *Aitchison* [1955].

Sample Sizes and Censoring

Of interest was the effect of censoring on data sets of varying sample sizes. Therefore three separate simulations were conducted, with data sets of 10, 25, 50 observations. In each simulation, 500 data sets were generated from each of the 16 parent distributions. All data sets were censored at four different levels (detection limits): the 20th, 40th, 60th, and 80th percentiles of the parent distributions. Such high percentages of censoring are common in trace level water quality data. With this "type I" censoring [*David*, 1981], the actual percentage of observations censored varied for each data set due to sample variability. For the gamma distribution with CV = 2.0, the 20th and 40th percentiles were so close to zero (0.0043 and 0.070) that they were discarded as being unrealistic detection limits.

We required the condition that at least three observations be present in each data set after censoring or the data set was discarded. For $n = 10$, this eliminated about 1% of data sets censored at the 40th percentile, about 18% at the 60th percentile, and about 72% at the 80th percentile. Results for censoring at the 80th percentile were therefore not considered meaningful for $n = 10$. For $n = 25$, less than 1% of data sets were eliminated at the 60th percentile censoring level and about 11% at the 80th percentile level. For $n = 50$, less than 1% of data sets censored at the 80th percentile were discarded.

PARAMETER ESTIMATION METHODS

There are many possible ways to estimate distributional parameters of censored data. Among the most commonly applied are ignoring censored observations, setting all censored

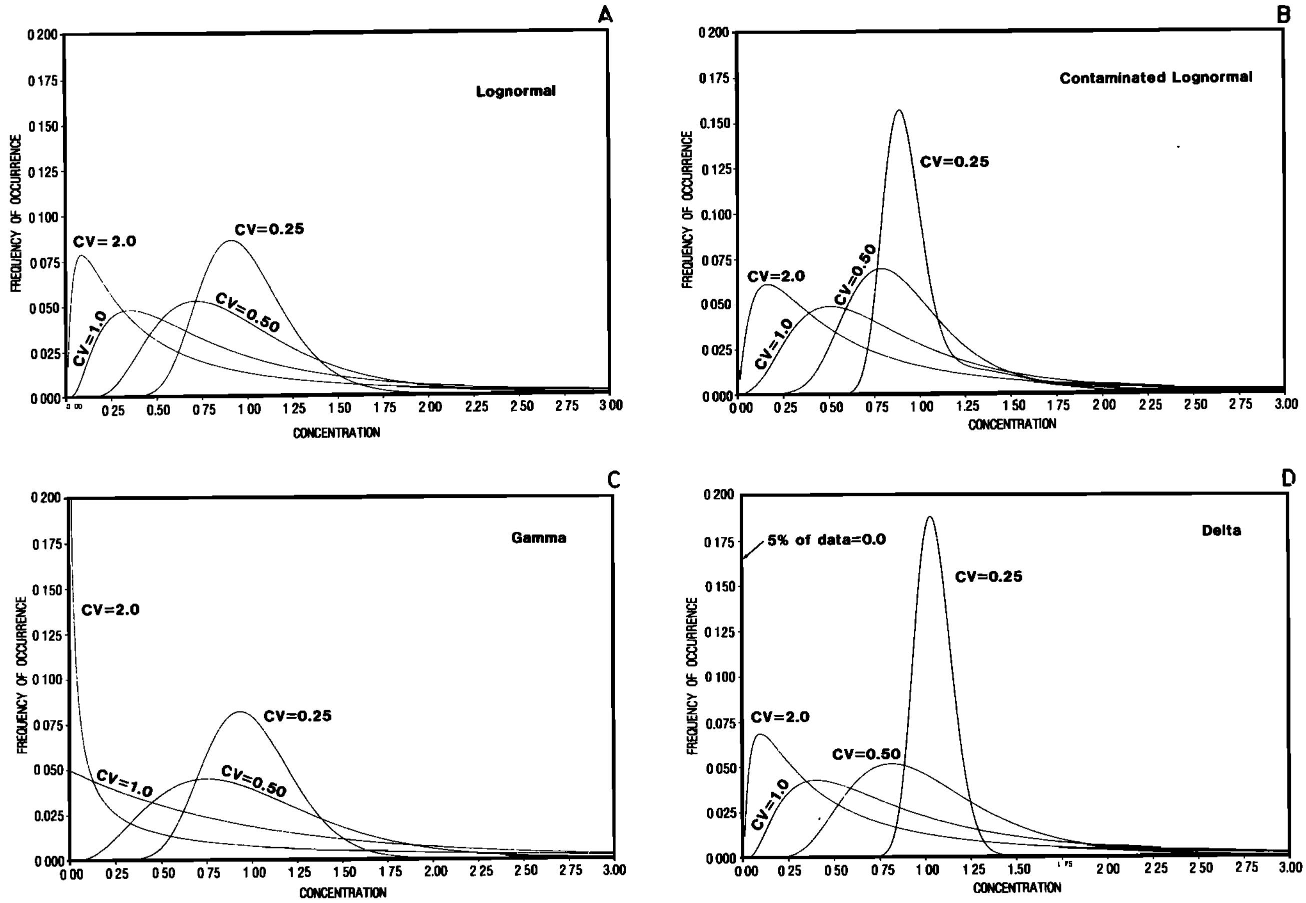


Fig. 1. Probability density functions for the parent distributions used in simulations.

observations equal to zero, or setting all censored observations equal to some fraction of the detection limit, and then using traditional computational methods. Another approach is to estimate the missing observations based on an assumed distribution of data between zero and the detection limit and then use traditional computational methods. Or, based on an assumption of the underlying distribution of the entire data set, maximum likelihood estimates of distributional parameters can be derived from the uncensored observations. In our experiments we evaluated eight methods for estimating the population mean, standard deviation, median, and interquartile range, representing all of these approaches. These are listed below along with their abbreviations used in this report.

1. ZE: censored observations were assumed to equal zero.

2. DL: censored observations were assumed to equal the detection limit.

3. UN: censored observations were assumed to follow a uniform distribution between zero and the detection limit. Thus for the ordered observations x_i, $i = 1, 2, \cdots, nc$ and nc = number of data censored, $x_i = \mathrm{dl}\,(i-1)/(nc-1)$, a distribution symmetric around one half the detection limit (dl).

4. NR: censored observations were assumed to follow the zero-to-detection limit portion of a normal distribution which was fit to the uncensored observations using least squares regression as follows. "Normal scores," z, were computed for each uncensored observation using

$$z = \Phi^{-1}(r/n + 1)$$

where Φ^{-1} is the inverse cumulative normal distribution function; r is the observation rank ($r = nc + 1, \cdots, n$); and n is the sample size for the entire data set. A least squares regression of concentration on normal scores for all data above the detection limit was extrapolated to estimate censored observations (ranks $r = 1, \cdots, nc$). Any estimated values falling below zero were set equal to zero.

5. LR: censored observations are assumed to follow the zero-to-detection limit portion of a lognormal distribution fit to the uncensored observations by least squares regression. The method is identical to NR, except that concentrations were log-transformed prior to analysis.

6. NM: concentrations are assumed to be normally distributed with parameters estimated from the uncensored observations by the maximum likelihood method for a censored normal distribution [*Cohen*, 1959].

7. LM: concentrations are assumed to be lognormally distributed with parameters estimated using logarithms of the uncensored observations in *Cohen*'s [1959] maximum likelihood method. The mean and standard deviation of the untransformed concentrations are then estimated using the equations given by *Aitchison and Brown* [1957].

8. DT: censored observations are assumed to be zero and uncensored observations are assumed to follow a lognormal distribution. Estimates of parameters of the overall delta distribution are obtained by computing maximum likelihood estimates of parameters of the uncensored lognormal portion and using relationships between these and the overall delta distribution described by *Aitchison* [1955].

The commonly used method of discarding censored observations prior to calculating parameter estimates was not included in this study. Discarding censored observations will always result in both higher bias and higher rmse than the DL method. Because this can never be the most appropriate (mini-

TABLE 1. Root Mean Squared Errors (rmses) of Estimation Methods for Data Sets of Size $n = 25$ in Percent of True Value

Mean		Standard Deviation		Median		Interquartile Range	
Method	rmse	Method	rmse	Method	rmse	Method	rmse
			Censored at 20th Percentile: 7500 Data Sets				
DL	20	UN	42	LM	16	LR	30
LR	20	NR	42	DT	19	LM	30
UN	21	LR	42	LR	19	DL	30
NR	21	DL	43	DL	19	NR	34
LM	21	NM	45	UN	19	UN	38
DT	22	ZE	58	NR	19	NM	52
NM	22	LM	76	ZE	19	ZE	138
ZE	23	DT	84	NM	41	DT	143
			Censored at 40th Percentile: 7500 Data Sets				
LR	20	LR	43	LM	17	LM	30
DL	21	NR	45	DL	18	LR	32
UN	22	DL	47	UN	19	DL	41
LM	22	UN	48	LR	20	NR	57
NR	23	NM	56	NR	30	UN	83
DT	31	ZE	76	ZE	45	NM	110
ZE	32	DT	90	DT	47	ZE	237
NM	42	LM	92	NM	52	DT	248
			Censored at 60th Percentile: 7994 Data Sets				
LR	23	LR	45	LM	63	LM	36
UN	25	NR	50	UN	75	LR	40
DL	29	UN	52	DL	87	DL	69
NR	29	DL	53	NR	90	NR	83
DT	45	ZE	80	LR	98	UN	121
ZE	45	NM	82	DT	107	NM	207
LM	79	DT	106	ZE	107	ZE	229
NM	104	LM	108	NM	403	DT	237
			Censored at 80th Percentile: 7148 Data Sets				
UN	29	LR	48	DT	100	LM	41
LR	30	UN	54	ZE	100	LR	45
LM	31	NR	55	NR	113	NR	94
NR	35	DL	63	LM	141	DL	96
DT	60	ZE	72	LR	201	UN	133
ZE	60	DT	118	UN	229	ZE	138
DL	61	NM	138	DL	369	DT	139
NM	224	LM	1300	NM	1000	NM	366

Methods are ranked by rmse.

mum rmse) method, it was not considered here. The commonly used substitution of values equal to one half the detection limit was also not included, due to its similarity to the UN method. These two methods will produce identical estimates for the mean, while a range in values between zero and the detection limit should produce better estimates of the other three parameters than substituting a single, arbitrary value for all censored data.

The evaluation of the reliability of estimation methods was based on rmses computed from actual parameters of the underlying distribution. Rmses for each parameter were computed for each estimation method and for each parent distribution. Deviations between the parameter values estimated from each censored data set and those of the underlying distribution were divided by the true population values to express rmses as fractions of the true values. For example, the equation for the rmse of the mean is

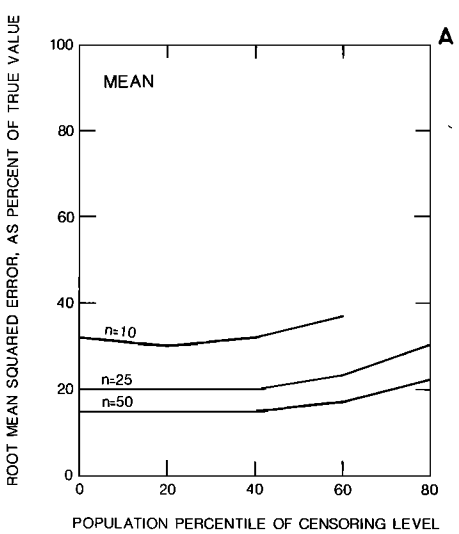
$$\text{rmse} = \left[\sum_{i=1}^{N} \left(\frac{\bar{x}_i - \mu}{\mu} \right)^2 \Big/ N \right]^{1/2} \tag{2}$$

where $\bar{x}_i$ is the estimate of the mean for the ith of N data sets. We also computed the bias portion of the rmse and the standard error of the rmse, which describes the reliability of our rmse estimates.

ESTIMATION WITHOUT CLASSIFICATION

Simulation results without classification of data sets are given in Table 1 for data sets of size $n = 25$ to show the typical pattern of results for all parameter estimation methods. Though rmses are higher and lower for $n = 10$ and $n = 50$, respectively, the same estimation methods always perform well for a particular combination of censoring level and distributional parameter.

There are several ways to approach identifying the "best" estimation method(s) from results such as those in Table 1. One approach would be to designate a best method for every single combination of censoring level, parameter, and sample size. Alternatively, a single robust method could be chosen that works well over the entire range of conditions simulated. Figure 2 illustrates these two method selection approaches. The best overall method was chosen by summing the ranks of rmses for each method over all sample sizes, censoring levels, and parameters. The method with the smallest sum of ranks, LR, was considered best. Rmses for LR are shown for all parameters in Figure 2, along with those for any other methods having rmses significantly ($\alpha = 0.05$) lower than that of LR. Little reduction in rmse for the mean and standard deviation is accomplished by considering different sample sizes and censoring levels separately. The rmses of LR are lowest, or not significantly different than the lowest, in virtually every situation.



EXPLANATION

n number of observations in each sample before censoring

—— RMSE of LR method

---- RMSE of best method (indicated for each datum)

Fig. 2. Root-mean-squared errors for best estimation methods.

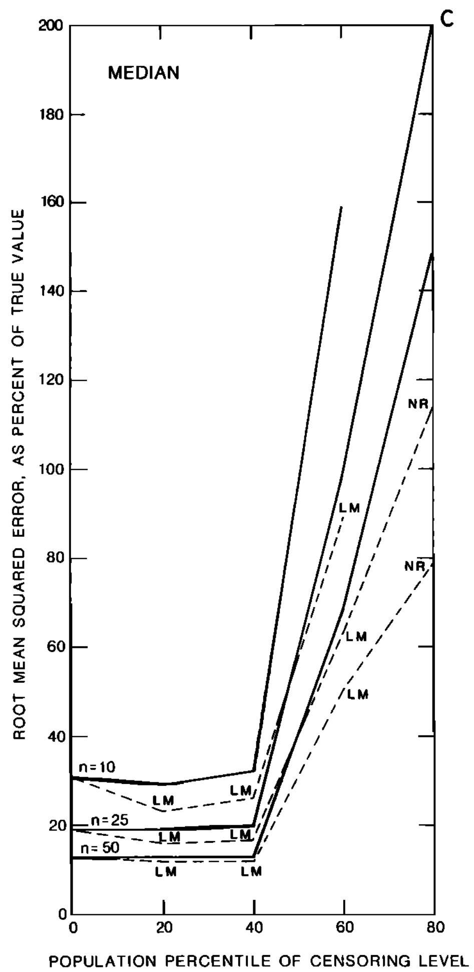


Fig. 2. (continued)

For the median and interquartile range, on the other hand, significant reductions in rmse can be achieved by using the best method for a particular set of conditions rather than using LR for all (Figure 2). The largest reductions in rmse occur for small sample sizes and high censoring. For all but four combinations of censoring level and sample size, the best method for estimating the median and interquartile range is LM. For the interquartile range at 20% censoring, LM is tied with LR for $n = 25$ and $n = 50$. For the median at 80% censoring and $n = 25$ and $n = 50$, LM is a close second to NR. For this latter case, DT and ZE results are ignored. These methods produced zero as the estimate of the median for every data set, merely an obvious lower bound. The resulting 100% bias and rmse are totally uninformative.

Figure 2, while showing the extremes of method selection approaches, suggests an effective third course: selecting LR for the mean and standard deviation and LM for the median and interquartile range. In fact, LR has the lowest sum of ranks (lowest rank with lowest rmse) of any method for the mean and standard deviation over all censoring levels and sample sizes while LM has the lowest sum of ranks for the median and interquartile range. Little reduction in rmse is accomplished by using other methods for differing sample sizes or censoring levels.

The LM method has been noted in Table 1 to produce some erratically high estimates of the mean and standard devi-

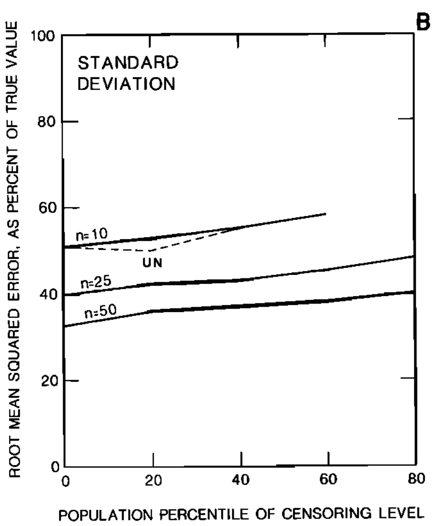


Fig. 2. (continued)

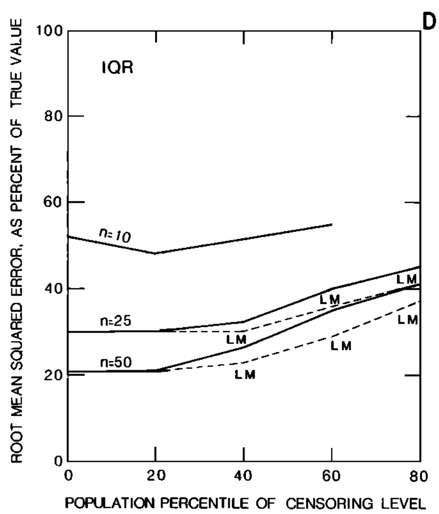


Fig. 2. (continued)

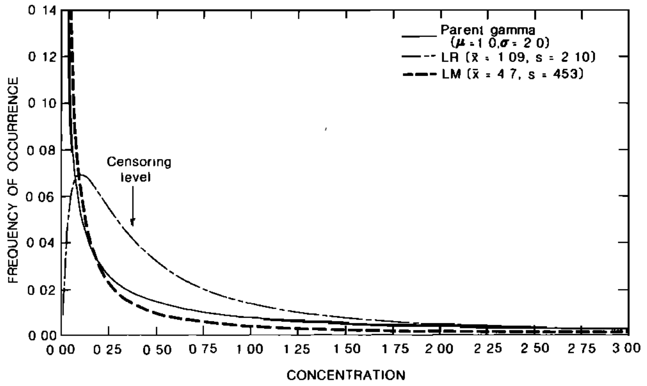


Fig. 3. Estimated frequency distributions by LM and LR for one data set ($n = 25$) from the GM(2.0) parent distribution compared to the parent distribution. Data set was censored at the 60th population percentile.

ation, particularly for higher censoring levels. This occurred for the same data sets for which LM generally produced the best estimates of the median and interquartile range. Figure 3 shows an example of the estimated probability distributions produced by the LM and LR methods, compared to the parent distribution for one data set generated from GM (2.0). The data set of 25 observations was censored at the 60th percentile. Figure 3 illustrates that the LM method produced an estimated distribution that more closely mimics the parent distribution than the LR method. This results in accurate estimates of percentiles. To do this, however, the mean and standard deviation were grossly overestimated at 4.7 and 453, respectively. The LR method, though not mimicking the shape of the parent distribution, produced accurate estimates of the mean (1.09) and standard deviation (2.10). Because the LR, NR, and UN methods involve simply calculating sample parameter statistics after estimating censored observations, they rarely produce wild estimates of distributional parameters.

The delta estimator (DT) was recommended by *Owen and DeRouen* [1980] for estimates of the mean in comparison to the LM method. However, their percent of data censored was known (type II censoring) and never exceeded 25%. With type I censoring at the 20th percentile, DT and LM give identical results (Table 1) for the mean, though not for other parameters or censoring levels. Both DT and LM are sensitive to extreme values at these small sample sizes, and therefore have higher errors than does LR.

ESTIMATION WITH CLASSIFICATION

Rankings and rmses were previously presented in Table 1 with all 16 parent distributions equally represented. If the parent distribution were known, however, the other 15 could be ignored, with the resulting method ranking and rmse mag-

TABLE 2. Rmses for Data Sets of Size $n = 25$ From Four Lognormal Parent Distributions Censored at the 80th Percentile in Percent of True Value

Mean		Standard Deviation		Median		Interquartile Range	
Method	rmse	Method	rmse	Method	rmse	Method	rmse
			LN (0.25) n = 443				
LM	9	LM	32	LM	11	LM	30
LR	12	LR	36	LR	15	LR	34
NM	17	NM	62	NM	17	NM	66
NR	22	DL	63	UN	22	NR	81
DL	23	NR	64	DL	23	DL	97
UN	24	UN	84	NR	26	UN	133
ZE	71	DT	97	ZE	100	ZE	168
DT	71	ZE	127	DT	100	DT	169
			LN (0.50) n = 450				
UN	13	UN	30	UN	10	LM	25
LM	14	LR	36	LM	21	LR	27
LR	20	LM	41	LR	29	UN	48
NR	33	NR	47	DL	49	NR	73
DL	43	ZE	55	NR	56	DL	97
ZE	64	DL	57	NM	65	ZE	112
DT	64	DT	90	ZE	100	DT	113
NM	64	NM	112	DT	100	NM	139
			LN (1.0) n = 458				
UN	20	UN	39	UN	33	UN	22
LM	22	ZE	42	LM	36	LM	29
LR	29	NR	44	LR	53	LR	32
NR	37	LR	47	NR	85	NR	72
ZE	52	DL	58	ZE	100	DL	95
DT	53	LM	75	DT	100	ZE	101
DL	67	DT	87	DL	101	DT	103
NM	178	NM	158	NM	225	NM	294
			LN (2.0) n = 457				
UN	39	ZE	53	LM	57	UN	29
LR	39	NR	53	LR	84	LM	40
NR	42	UN	56	UN	90	LR	43
LM	47	LR	57	NR	100	NR	84
DT	48	DL	65	ZE	100	DL	94
ZE	49	DT	125	DT	100	ZE	101
DL	77	NM	156	DL	191	DT	103
NM	366	LM	866	NM	734	NM	620

Methods are ranked by rmse.

TABLE 3. Groups of Parent Distributions for Which Best Performing Methods and Their rmses Were Similar

Population Percentile of Censoring Level	Group I	Group II	Group III	Group IV	Group V	Group VI
20	LN(0.25)	LN(0.50)	LN(1.0)	LN(2.0)		
	GM(0.25)	GM(0.50)	GM(1.0)	DT(2.0)		
	DT(0.25)	DT(0.50)	DT(1.0)	CT(2.0)		
	CT(0.25)	CT(0.50)	CT(1.0)			
40	LN(0.25)	LN(0.50)	LN(1.0)	LN(2.0)		
	GM(0.25)	GM(0.50)	GM(1.0)	DT(2.0)		
	DT(0.25)	DT(0.50)	DT(1.0)	CT(2.0)		
	CT(0.25)	CT(0.50)	CT(1.0)			
60	LN(0.25)	CT(0.25)	LN(0.50)	LN(1.0)	LN(2.0)	GM(2.0)
	GM(0.25)		GM(0.50)	GM(1.0)	DT(2.0)	
	DT(0.25)		DT(0.50)	DT(1.0)	CT(2.0)	
			CT(0.50)	CT(1.0)		
80	LN(0.25)	CT(0.25)	LN(0.50)	LN(1.0)	LN(2.0)	GM(2.0)
	GM(0.25)		GM(0.50)	GM(1.0)	DT(2.0)	
	DT(0.25)		DT(0.50)	DT(1.0)	CT(2.0)	
			CT(0.50)	CT(1.0)		

LN, lognormal; CT, contaminated lognormal; DT, delta; GM, gamma.

nitudes possibly quite different than Table 1. For example, Table 2 separately presents rmses for data sets from each of the four lognormal distributions. All data sets consisted of 25 observations and were censored at the 80th percentile. For a lognormal distribution with CV = 0.25, the lowest ranked estimation method (LM) for the mean has an rmse of 9%, while for CV = 2.0 it is either the UN or LR methods with an rmse of 39% (Table 2). Table 1, on the other hand, shows that over all 16 distributions the UN method is ranked lowest for estimating the mean, with an rmse of 29%. Therefore if the parent distribution of a data set could be inferred from attributes of data above the detection limit, improved method selection and estimates of rmse magnitude should result. This is the goal of classification.

Note that if the true distribution were LN (2.0), the rmse of 39% would be greater than that estimated in Table 1, and yet would be more accurate, because Table 1 incorporates rmses from the lower error distributions.

Selection of Class Boundaries

To define class boundaries for estimation method selection, the following procedure was repeated for each of the four censoring levels.

1. The performance of parameter estimation methods was evaluated separately for data sets ($n = 25$) from each of the 16 parent distributions at each censoring level. For each censoring level, individual parent distributions with similar best performing estimation methods and similar rmses were grouped together. These groups of similar distributions, which reflect the dominant effect of population coefficient of variation on estimation error, are given in Table 3.

2. Four dimensionless sample statistics were computed from the data above the detection limit for all simulated data sets. These sample statistics were

Coefficient of skewness

$$g = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x}_u)^3}{s_u^{\,3}}$$

Coefficient of variation

$$\mathrm{CV} = \frac{s_u}{\bar{x}_u}$$

Quartile estimate of skew

$$qs = \frac{q_3 - 2q_2 + q_1}{q_3 - q_1}$$

Relative quartile range

$$rqr = \frac{q_3 - q_1}{d}$$

where

- k number of uncensored observations;
- x_i individual observation in data set;
- $\bar{x}_u$ sample mean of uncensored observations;
- s_u sample standard deviation of uncensored observations;
- q_1, q_2, q_3 25th, 50th, and 75th sample percentiles of uncensored observations;
- d detection limit

3. The effectiveness of these four statistics for classifying each data set into the correct group of parent distributions was evaluated using box plots of the distribution of each sample statistic for each group. The most effective statistic was the relative quartile range (rqr), a measure of the dispersion of data above the detection limit relative to the magnitude of the detection limit. Box plots of rqr for data sets from each group of parent distributions are shown in Figure 4.

4. The best separation between groups, based on rqr at sample size 50, was evaluated using pairwise discriminant analysis. A lognormal distribution of rqr's was assumed, due to the asymmetry of the box plots, and the probability density function equations for each consecutive group pair were solved. The point at which two densities were equal was the optimum point of separation. Each density was weighted by the number of data sets per group. When no solution occurred, the two groups could not be distinguished by rqr (for example, groups II and III for censoring at the 80th percentile). The resulting class boundaries are also shown in Figure 4. Note that, since some distribution groups could not be discriminated, some rqr classes represent two predominant distribution groups.

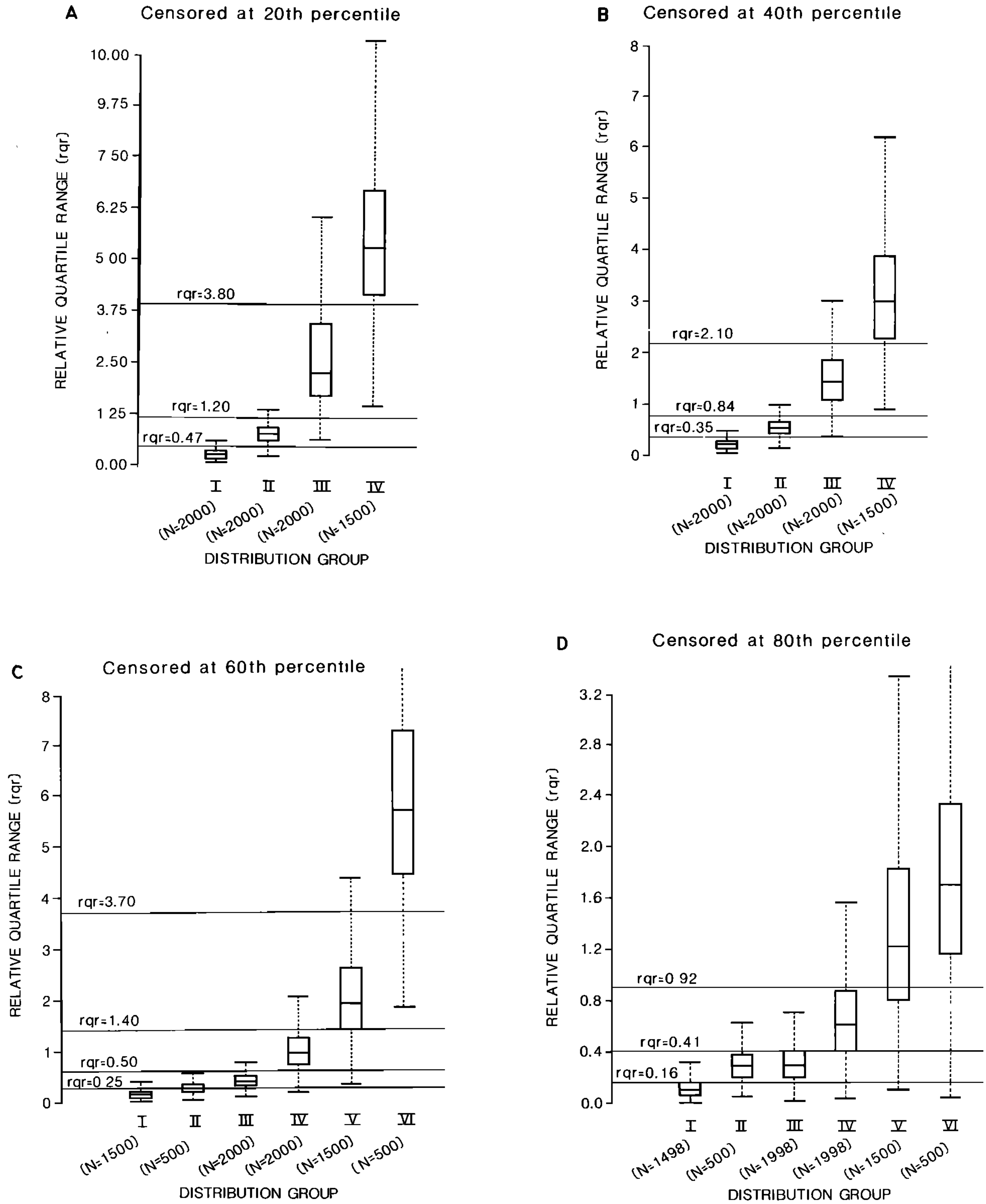


Fig. 4. Box plots of the relative quartile range for N data sets (Sample size = 50) from each group of parent distributions (Table 3) and class boundaries determined by discriminant analysis.

Benefits of Classification

The 500 data sets for each of the 16 parent distributions were censored at the four levels, and then classified using the class boundaries developed by discriminant analysis. Figure 5 shows the success of classifying data sets into the group containing their parent distribution. A decrease in classification success with decreasing sample size and increasing censoring level is evident. This reflects the smaller amount of information contained in small data sets and the loss of information due to censoring. The class boundaries determined by discriminant analysis of *rqr* for data sets of 50 observations (and shown in Figure 4) are superior or equal to those determined from data sets of 25 observations, with only one exception. This is not surprising, as more information is present at

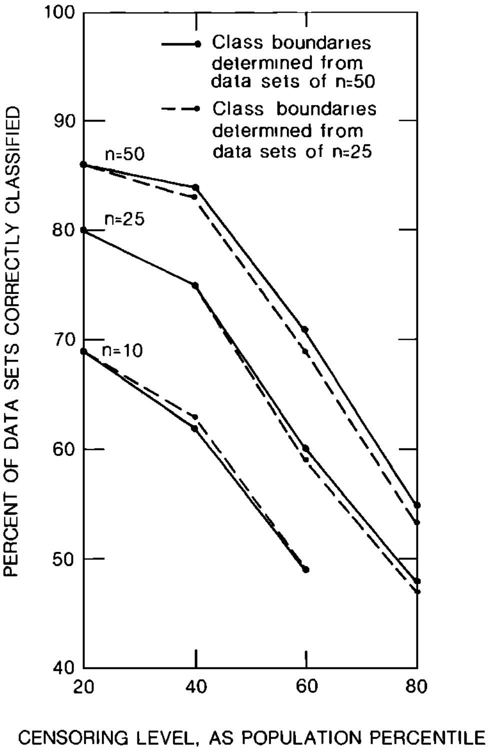


Fig. 5. Success of data set classification.

the larger sample size. Class boundaries from data sets of 10 observations were much less effective than either of the two shown in Figure 5. Therefore the boundaries determined from 50-observation data sets were used in all subsequent classifications.

METHOD SELECTION

The best estimation method was determined for each combination of sample size, censoring level, and *rqr* class. In light of the results without classification, best methods for the mean and standard deviation were determined separately from those for the median and interquartile range. The best method was that which minimized the ranks of rmses across the two distributional parameters being considered. If additional methods had rmses not significantly different (t test at $\alpha = 0.05$) from the best for both parameters, these were also included as "best." Finally, a single best method over all three sample sizes was selected for each *rqr* class; results are given in Table 4. The single best method was often the only method that qualified for best for all three sample sizes. Where more than one method qualified or where none was best over all sample sizes, the method which minimized the sum of squared rmses over the three sample sizes was selected.

The classification system shown in Table 4 sometimes results in different method selection than that obtained without classification and shown in table 1. The LM method remains the best method for the median and interquartile range for all *rqr* classes. However, whereas results without classification indicated that the LR method was generally best for the mean and standard deviation, results in Table 4 show that these distributional parameters are often best estimated by the LM, UN, or NR methods.

Table 5 compares rmses of the best methods for $\bar{x}$ and s for each *rqr* class (from Table 4) to the corresponding rmses of LR, the best overall method without classification. This comparison shows that in most instances there is no significant difference ($\alpha = 0.05$) between the rmse of LR compared to the rmse of the best method chosen according to the criteria described. Even where differences are statistically significant, they are not large. In contrast, neither LM, UN, nor NR are

TABLE 4. Rmses of Best Estimation Methods When Classified by *Rqr* in Percent of True Value

	Censored at 20th Percentile				Censored at 40th Percentile				Censored at 60th Percentile				Censored at 80th Percentile			
	$\bar{x}$	s	m	iqr	$\bar{x}$	s	m	iqr	$\bar{x}$	s	m	iqr	$\bar{x}$	s	m	$iqr.$
	Rqr < 0.47				*Rqr* < 0.35				*Rqr* < 0.25				*Rqr* < 0.16			
Best methods	LM		LM		LM		LM		LM		LM		LM		LM	
$n = 10$	11	43	10	37	13	65	9	39	10	51	14	42				
$n = 25$	6	35	5	26	7	36	6	27	8	40	8	30	12	48	22	39
$n = 50$	4	32	4	18	5	33	4	23	6	36	5	26	9	44	22	32
	Rqr = 0.47–1.2				*Rqr* = 0.35–0.84				*Rqr* = 0.25–0.60				*Rqr* = 0.16–0.41			
Best methods	LM		LM		LM		LM		LM		LM		LR		LM	
$n = 10$	19	40	17	38	20	44	18	39	18	45	27	43				
$n = 25$	11	25	10	25	12	28	11	26	14	33	15	35	24	42	82	44
$n = 50$	7	19	7	18	8	20	8	20	9	23	10	34	18	31	39	44
	Rqr = 1.2–3.8				*Rqr* = 0.84–2.1				*Rqr* = 0.6–1.4				*Rqr* = 0.41–0.92			
Best methods	UN		LM		UN		LM		UN		LM		UN		LM	
$n = 10$	32	56	26	45	29	53	28	44	26	52	63	46				
$n = 25$	23	47	19	31	22	43	20	29	21	46	27	33	20	46	150	43
$n = 50$	16	36	13	23	15	35	14	21	15	35	17	23	17	42	94	38
	Rqr > 3.8				*Rqr* > 2.1				*Rqr* = 1.4–3.7				*Rqr* > 0.92			
Best methods	NR		LM		NR		LM		UN		LM		LR		LM	
$n = 10$	54	65	37	77	60	73	46	81	47	64	130	70				
$n = 25$	34	52	25	40	35	57	27	41	31	54	77	38	44	56	240	37
$n = 50$	25	49	18	26	26	51	19	28	25	50	46	28	32	48	200	28
									Rqr > 3.7							
Best methods									NR		LM					
$n = 10$									85	94	240	110				
$n = 25$									45	53	200	57				
$n = 50$									30	43	170	39				

TABLE 5. Rmses of Best Method Compared to rmses of LR for the Mean and Standard Deviation in Each *rqr* Class

	Censored at 20th Percentile		Censored at 40th Percentile		Censored at 60th Percentile		Censored at 80th Percentile	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
	Rqr < 0.47		*Rqr* < 0.35		*Rqr* < 0.25		*Rqr* < 0.16	
Method	LM/LR		LM/LR		LM/LR		LM/LR	
$n = 10$	11/11	*43/48	13/13	65/52	*10/13	*51/56		
$n = 25$	6/6	*35/37	7/7	*36/39	*8/9	*40/43	*12/19	48/50
$n = 50$	4/4	32/33	5/5	33/35	6/7	36/37	*9/12	44/43
	Rqr = 0.47–1.2		*Rqr* = 0.35–0.84		*Rqr* = 0.25–0.60		*Rqr* = 0.16–0.41	
Method	LM/LR		LM/LR		LM/LR		LR	
$n = 10$	19/20	40/43	20/21	44/45	*18/21	45/47		
$n = 25$	11/11	*25/32	12/12	*28/34	*14/16	*33/38	24	42
$n = 50$	7/7	*19/22	8/8	*20/25	*9/10	*23/29	18	31
	Rqr = 1.2–3.8		*Rqr* = 0.84–2.1		*Rqr* = 0.60–1.4		*Rqr* = 0.41–0.92	
Method	UN/LR		UN/LR		UN/LR		UN/LR	
$n = 10$	32/32	56/57	29/30	50/55	*26/31	52/53		
$n = 25$	23/22	47/47	22/21	43/44	21/23	46/46	*20/29	46/45
$n = 50$	16/16	36/37	15/15	35/36	15/15	35/36	*17/22	42/40
	Rqr > 3.8		*Rqr* > 2.1		*Rqr* = 1.4–3.7		*Rqr* > 0.92	
Method	NR/LR		NR/LR		UN/LR		LR	
$n = 10$	54/55	65/65	60/62	73/73	47/51	64/65		
$n = 25$	34/34	52/53	35/35	57/57	31/31	54/54	44	56
$n = 50$	25/25	49/49	26/26	51/51	25/25	50/50	32	48
					Rqr > 3.7			
Method					NR/LR			
$n = 10$					85/93	94/93		
$n = 25$					45/50	53/53		
$n = 50$					30/33	43/43		

*Significant difference at $\alpha = 0.05$.

similarly robust over all *rqr* classes. For example, Table 5 indicates that LM has a significantly lower rmse than LR for both the mean and standard deviation at the 60th percentile censoring level and *rqr* = 0.25–0.60 (n = 25). Yet LM is the worst method in the next highest *rqr* class (*rqr* = 0.60–1.4) for both the mean and standard deviation, with rmses over 100% of the true value for standard deviation.

When applying parameter estimation methods to actual water quality data, an important consideration is method robustness. Given the possibility of misclassifying individual

TABLE 6. Rmses When All Data Sets (n = 25) are Classified Correctly by Distribution Group (Perfect) as Compared to Results of Actual Classification from Tables 4 and 5

	Censored at 20th Percentile				Censored at 40th Percentile				Censored at 60th Percentile				Censored at 80th Percentile			
	$\bar{x}$	s	m	*iqr*	$\bar{x}$	s	m	*iqr*	$\bar{x}$	s	m	*iqr*	$\bar{x}$	s	m	*iqr*
	Rqr < 0.47				*Rqr* < 0.35				*Rqr* < 0.25				*Rqr* < 0.16			
Method	LR		LM		LR		LM		LR		LM		LR		LM	
Perfect	5	36	4	25	6	38	4	28	7	41	6	26	5	30	4	29
Actual	6	37	3	26	7	39	6	27	9	43	8	30	19	50	22	39
	Rqr = 0.47–1.2				*Rqr* = 0.35–0.84				*Rqr* = 0.25–0.60				*Rqr* = 0.16–0.41			
Method	LR		LM		LR		LM		LR		LM		LR		LM	
Perfect	10	29	10	25	10	31	11	25	14	33	14	27	22	39	23	29
Actual	11	32	10	25	12	34	11	26	16	38	15	35	24	42	82	44
	Rqr = 1.2–3.8				*Rqr* = 0.84–2.1				*Rqr* = 0.60–1.4				*Rqr* = 0.41–0.92			
Method	LR		LM		LR		LM		LR		LM		LR		LM	
Perfect	20	47	19	30	20	44	20	30	22	46	23	30	30	48	38	24
Actual	22	47	19	31	21	44	20	29	23	46	27	33	29	45	150	43
	Rqr > 3.8				*Rqr* > 2.1				*Rqr* = 1.4–3.7				*Rqr* > 0.92			
Method	LR		LM		LR		LM		LR		LM		LR		LM	
Perfect	36	60	25	41	36	60	27	41	36	61	33	40	39	61	55	35
Actual	34	53	25	40	35	57	27	41	31	54	77	38	44	56	240	37
									Rqr > 3.7							
Method									LR		LM					
Perfect									58	43	240	41				
Actual									50	53	200	57				

Rmses are in percent of true value.

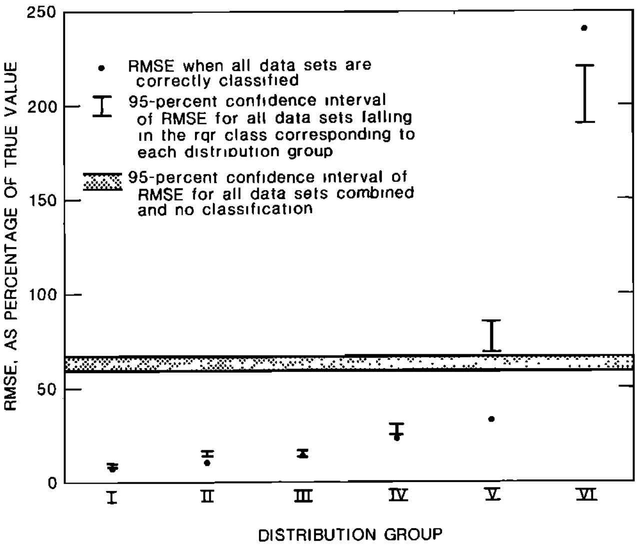


Fig. 6. Comparison of rmses with and without classification for estimates of the median from data sets of $n = 25$ censored at the 60th population percentile.

data sets (Figure 5), and the small increases in rmse when LR is used for any *rqr* class, the use of the more robust LR method is best for making low-risk estimates of the mean and standard deviation for all data sets.

ACCURACY OF RMSES

Though the classification system does not, in practice, alter method selection compared to results with no classification, it does result in superior estimates of error (rmse), by considering differences due to the probable parent distribution. Table 2 showed that rmses vary considerably between data sets from different parent distributions. The classification system was designed to indicate the types of parent distributions from which each data set may have originated, and therefore yield more accurate estimates of error (whether higher or lower) than the average rmse for all data sets from all 16 parent distributions, such as given in Table 1.

Table 6 shows rmses for the best parameter estimation methods (LR for $\bar{x}$ and s, LM for m and *iqr*) for data sets only from parent distributions intended to be included in each *rqr* class (Table 2 and Figure 4). These rmses represent the reliability of parameter estimates if each data set were correctly classified according to its parent distribution. Also shown in Table 6 for comparison are the previously reported rmses for data sets actually falling in each *rqr* class during the simulation with all 16 parent distributions (from Table 5 for $\bar{x}$ and s by the LR method, and Table 4 for m and *iqr* by the LM method).

Table 6 shows that the *rqr* classification system results in rmses which are very similar to the best estimate of true rmse, that of perfect classification. Only at 80th percentile censoring do the rmse values substantially depart from truth. This reflects the greater inability to correctly classify highly censored data sets previously illustrated in Figure 5. Even at 80th percentile censoring, however, *rqr* classification generally improves the accuracy of rmse estimates over those with no classification.

To illustrate the improvement in rmse accuracy following classification, the data for 60th percentile censoring ($n = 25$) is plotted in Figure 6. Shown in the figure are the rmses for perfect classification into parent distribution group, those for the actual classification according to *rqr*, and the rmse without classification. When data sets are classified, more reliable rmse estimates are obtained.

CONCLUSIONS

The most robust estimation method for minimizing errors in estimates of the mean, standard deviation, median, and interquartile range of censored data was the log probability regression method (LR). This method is based on the assumption that censored observations follow the zero-to-censoring level portion of a lognormal distribution obtained by a least squares regression between logarithms of uncensored concentration observations and their normal scores.

When method performance was evaluated separately for each distributional parameter, LR resulted in the lowest rmses for the mean and standard deviation. The lognormal maximum likelihood estimator for censored data (LM) produced lowest rmses for the median and interquartile range. These two methods constitute the best procedures for their respective parameters.

Using the relative quartile range (*rqr*), the interquartile range of uncensored observations divided by the detection limit, censored data sets can be classified into groups reflecting their probable parent distribution. Within these *rqr* groups, the accuracy of rmses substantially improved over those without classification.

These findings appear to have great potential for improving estimation of distributional parameters from censored water quality data sets. However, to apply the results of these Monte Carlo experiments to censored trace water-quality data, several assumptions are required. In addition, the *rqr* classification system and rmses need to be verified with actual water quality data sets. These issues are addressed in detail in a companion paper [*Helsel and Gilliom*, this issue].

APPENDIX: EQUATIONS FOR THE CONTAMINATED LOGNORMAL DISTRIBUTION

$$\mu_x = (1 - p)\mu_{x1} + p\mu_{x2} \tag{A1}$$

$$\sigma_x = \mu_x \left[\frac{C_3 (\sigma_{x1}/\mu_{x1})^2 + C_2}{C_1}\right]^{1/2} \tag{A2}$$

where

$$C_1 = (1 - p + pk)^2; \tag{A3}$$

$$C_2 = p(1 - p)(1 - k)^2; \tag{A4}$$

$$C_3 = 1 - p + 4pk^2; \tag{A5}$$

p percent of population described by the lognormal distribution with μ_{x2} and σ_{x2};

k ratio of $\mu_{x2} : \mu_{x1}$.

Algebraic manipulation of (A1)–(A5) leads to the following relationships for the two individual distributions which make up the overall contaminated lognormal distribution:

$$\mu_{x1} = \frac{\mu_x}{(1 - p + pk)} \tag{A6}$$

$$\sigma_{x1} = \mu_{x1}\left[\left(\frac{\sigma_x}{\mu_x}\right)^2 - \frac{C_2}{C_1}\right]^{1/2}\left(\frac{C_1}{C_3}\right)^{1/2} \tag{A7}$$

$$\mu_{x2} = \mu_{x1}k \tag{A8}$$

$$\sigma_{x2} = 2\mu_{x2} \frac{\sigma_{x1}}{\mu_{x1}} \tag{A9}$$

Given the specified conditions of the Monte Carlo simulation (μ_x and σ_x/μ_x), (A6)–(A9) yield estimates of μ_{x1}, σ_{x1}, μ_{x2}, and σ_{x2} which are used to generate two lognormal distributions. To generate data sets from the overall distribution, 80% of each data set was generated according to μ_{x1}, σ_{x1}, and 20% according to μ_{x2}, σ_{x2}.

Acknowledgments. The authors are deeply indebted to Jerry R. Stedinger and Robert M. Hirsch for many fruitful and insightful suggestions.

REFERENCES

Aitchison, J., On the distribution of a positive random variable having a discrete probability mass at the origin, *J. Am. Stat. Assoc.*, *50*, 901–908, 1955.

Aitchison, J., and J. A. C. Brown, *The Lognormal Distribution*, 176 pp., Cambridge, University Press, New York, 1957.

Cohen, A. C., Jr., Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics*, *1*(3), 217–237, 1959.

Condie, R., and K. A. Lee, Flood frequency analysis with historic information, *J. Hydrol.*, *58*, 47–61, 1982.

David, H. A., *Order Statistics, 2nd ed.*, 360 pp., John Wiley, New York, 1981.

Hashimoto, L. K., and R. R. Trussell, Evaluating water quality data near the detection limit, paper presented at the Proceedings of the American Water Works Assoc. Advanced Technology Conference, Am. Water Works Assoc., Las Vegas, Nev., June 5–9, 1983.

Helsel, D. R., and R. J. Gilliom, Estimation of distributional parameters for censored trace level water quality data, 2, Verification and applications, *Water Resour. Res.*, this issue.

Leese, M. N., Use of censored data in the estimation of Gumbel distribution parameters for annual maximum flood series, *Water Resour. Res.*, *9*(6), 1534–1542, 1973.

Owen, W. J., and T. A. DeRouen, Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants, *Biometrics*, *36*, 707–719, 1981.

R. J. Gilliom and D. R. Helsel, U.S. Geological Survey, 410 National Center, Reston, VA 22092.

(Received January 22, 1985;
revised September 10, 1985;
accepted October 23, 1985.)