WATER RESOURCES RESEARCH, VOL. 22, NO. 2, PAGES 147–155, FEBRUARY 1986

Estimation of Distributional Parameters for Censored Trace Level Water Quality Data 2. Verification and Applications

DENNIS R. HELSEL AND ROBERT J. GILLIOM

U.S. Geological Survey, Reston, Virginia

Estimates of distributional parameters (mean, standard deviation, median, interquartile range) are often desired for data sets containing censored observations. Eight methods for estimating these parameters have been evaluated by R. J. Gilliom and D. R. Helsel (this issue) using Monte Carlo simulations. To verify those findings, the same methods are now applied to actual water quality data. The best method (lowest root-mean-squared error (rmse)) over all parameters, sample sizes, and censoring levels is log probability regression (LR), the method found best in the Monte Carlo simulations. Best methods for estimating moment or percentile parameters separately are also identical to the simulations. Reliability of these estimates can be expressed as confidence intervals using rmse and bias values taken from the simulation results. Finally, a new simulation study shows that best methods for estimating uncensored sample statistics from censored data sets are identical to those for estimating population parameters. Thus this study and the companion study by Gilliom and Helsel form the basis for making the best possible estimates of either population parameters or sample statistics from censored water quality data, and for assessments of their reliability.

INTRODUCTION

Water quality data often include observations measured only as less than the detection limit, resulting in censored data sets. Eight methods for estimating distributional parameters for censored water quality data were evaluated by *Gilliom and Helsel* [this issue]. Results of extensive Monte Carlo simulations, in which large numbers of small samples were generated from 16 different parent distributions and censored to varying degrees, indicated that a log-probability regression method (LR) was the best method for estimating the mean and standard deviation of censored data and that a lognormal maximum likelihood method (LM) was best for estimating the median and interquartile range. That study, hereafter called the simulation study, also showed that censored data sets could be effectively classified using a sample statistic called the relative quartile range (*rqr*), which is the interquartile range of uncensored observations divided by the detection limit. Classification of simulation data sets according to *rqr* indicated the probable underlying distribution, and resulted in improved estimates of the precision of distributional parameters as compared to unclassified data sets.

The purposes of this study are to (1) verify the findings from the previous simulation study by evaluating the same parameter estimation methods using actual water-quality data; (2) describe an approach for estimating confidence bounds around parameter estimates made from censored water quality data; and (3) evaluate how well the estimation methods calculate uncensored sample statistics from censored data sets and compare their errors to those for estimating population parameters.

VERIFICATION OF PREVIOUS SIMULATION STUDY

Evaluations of parameter estimation methods in the previous simulation study are verified by applying the same type of analysis to actual water quality data. The best performing parameter estimation methods for actual water quality data are compared to the simulation study results. The *rqr* classification system developed in the simulation study is tested by comparing method performance for actual and simulated data within each *rqr* class, and by evaluating the ability of *rqr* classification to separate water quality data sets having different root mean squared errors (rmses) of parameter estimates.

Approach

Uncensored data sets with more than 50 observations for suspended sediment, total phosphorus, total Kjeldahl nitrogen, and nitrate nitrogen concentrations were obtained from 313 stations of the U.S. Geological Survey's National Stream Quality Accounting Network (NASQAN). Most data were monthly samples taken during 1974–1981, resulting in 917 data sets having more than 50 observations and no censoring.

Suspended sediment and major nutrients data were analyzed rather than trace constituents because (1) most available data sets for trace constituents consisted of less than 30 observations; (2) most trace constituent data sets contained censored observations; and (3) suspended sediment and nutrients are transported by the same types of processes as many trace constituents.

This last point is important because similarity in transport process will tend to result in similarly shaped frequency distributions. We examined this assumption by comparing the distributions of coefficients of variation (CV) and of a measure of symmetry between subsamples of $n = 25$ from each of the sediment and nutrient data sets and uncensored trace-constituent data sets of sizes ranging from $n = 20$ to $n = 40$. The measure of symmetry, *ms*, was

$$ms = \frac{q_{75} - q_{50}}{q_{50} - q_{25}} \quad (1)$$

where q_i is the ith percentile of the data set. The results of the comparison are shown in Figure 1, which also includes the same information for simulation study data sets (100 data sets from each of the 16 parent distributions) of size $n = 25$. All three types of data have similar distributions of these nondimensional variance and symmetry sample statistics.

For the verification tests, two subsamples, one of size $n = 10$ and one of $n = 25$, were randomly selected with replacement

Paper number 5W4031.

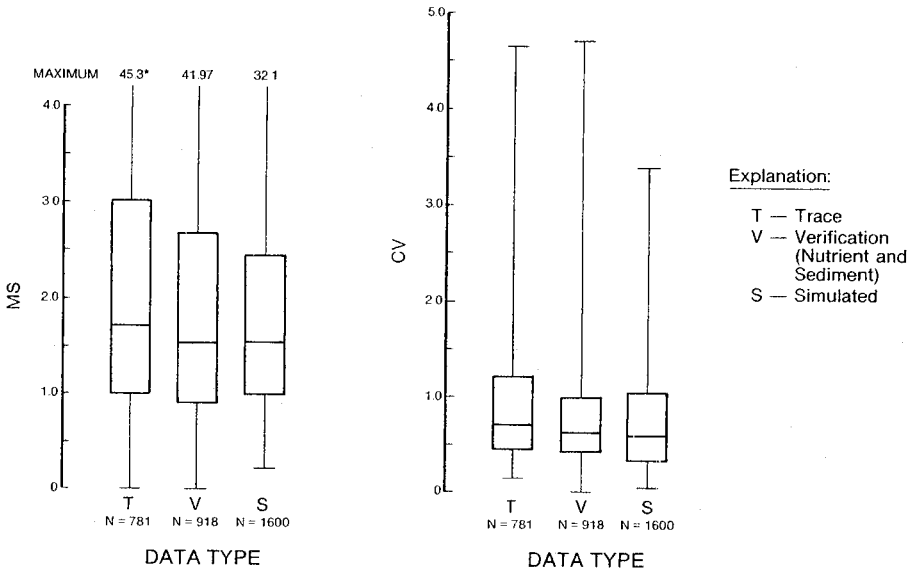

Fig. 1. Symmetry measure (MS) and coefficient of variation (CV) for three data types (35 data sets have denominator = 0, and are beyond the "maximum").

from each of the 917 sediment and nutrient data sets. Each resulting small sample was censored at 20, 40, 60, and 80% by the type II method [*David*, 1981], as population percentiles were not known. With this method the same fraction of each data set is censored. Each of the eight parameter estimation methods evaluated in the simulation study (Table 1) were applied to each censored sample. Rmses were computed for the mean, standard deviation, median, and interquartile range. Sample statistics computed from the original ($n > 50$) sediment and nutrient data sets were used as estimates of the true population parameters in rmse calculations.

Results

Best methods for the verification data, methods with the lowest rmse or with rmses not significantly (t test at $\alpha = 0.05$) larger than the lowest, were identical to those of the simulation study. Table 2 presents rmses for data sets of $n = 25$. Similar ordering of methods, though with higher rmses, were found for $n = 10$.

The best overall method for estimating the mean, standard deviation, median, and interquartile range of simulated data

TABLE 1. Parameter Estimation Methods

Method	Description of Method
ZE	Censored observations set to zero.
DL	Censored observations set to the detection limit.
UN	Censored observations uniformly distributed between zero and the detection limit.
NR	Censored observations followed the zero-to-detection limit portion of a normal distribution fit to uncensored observations by least squares regression.
LR	Censored observations followed the zero-to-detection limit portion of a lognormal distribution fit to uncensored observations by least squares regression.
NM	Maximum likelihood method for censored normal distributions.
LM	Maximum likelihood method for censored normal distributions using natural logarithms, followed by *Aitchison and Brown* [1957] transformation.
DT	Delta distribution estimator of *Aitchison* [1955].

See *Gilliom and Helsel* [this issue] for further detail.

TABLE 2. Root Mean Squared Errors (rmses) of Estimation Methods for 917 Verification Data Sets of Size $n = 25$ in Percent of Uncensored Value

Mean		Standard Deviation		Median		Interquartile Range	
Method	Rmse	Method	Rmse	Method	Rmse	Method	Rmse
			20% Censored				
NM	23	UN	41	LR	20	LR	47
DT	25	NR	41	UN	20	UN	47
LR	26	LR	42	ZE	20	ZE	47
DL	26	DL	42	DL	20	DL	47
UN	26	ZE	45	NR	20	NM	47
NR	26	NM	45	LM	20	LM	47
ZE	27	LM	84	DT	25	NR	47
LM	33	DT	*	NM	173	DT	59
			40% Censored				
LR	26	UN	42	LM	18	LM	50
UN	27	LR	43	LR	20	DL	51
DL	28	NR	44	UN	20	LR	52
NR	28	DL	46	ZE	20	UN	53
DT	32	ZE	55	DL	20	NR	80
ZE	33	NM	56	NR	20	ZE	129
NM	47	LM	*	DT	21	DT	146
LM	51	DT	*	NM	527	NM	757
			60% Censored				
LR	27	UN	44	LM	27	LM	54
UN	29	LR	45	UN	30	LR	56
NR	32	NR	47	LR	37	UN	65
DL	35	DL	52	DL	41	DL	73
ZE	44	ZE	60	NR	64	NR	91
DT	44	NM	79	ZE	100	ZE	129
NM	111	DT	*	DT	100	DT	144
LM	*	LM	*	NM	*	NM	*
			80% Censored				
UN	36	UN	44	NR	80	LM	72
LR	37	LR	47	ZE	100	UN	85
NR	43	NR	50	DT	100	LR	97
DT	62	ZE	54	LM	124	ZE	100
ZE	62	DL	63	UN	217	DL	100
DL	68	NM	133	LR	262	DT	100
LM	81	DT	*	DL	358	NR	125
NM	296	LM	*	NM	*	NM	*

*Rmse $\geq$ 1000%.

TABLE 3. Comparison of Rmses From Simulation Results to rmses From Verification Results

Method	Censored at 20th Percentile					Censored at 40th Percentile					Censored at 60th Percentile					Censored at 80th Percentile				
	LR			LM		LR			LM		LR			LM		LR			LM	
	N	$\bar{x}$	s	m	iqr	N	$\bar{x}$	s	m	iqr	N	$\bar{x}$	s	m	iqr	N	$\bar{x}$	s	m	iqr
	Rqr < 0.47					*Rqr < 0.35*					*Rqr < 0.25*					*Rqr < 0.16*				
$n = 10$	127	*11/24	*48/59	*10/16	*37/43	187	*13/26	*52/60	*9/20	39/43	176	*13/30	*56/67	*14/41	*42/52					
$n = 25$	104	*6/13	37/42	*5/7	26/27	160	*7/16	39/44	*6/9	27/30	176	*9/15	43/47	*8/26	*30/38	297	*19/35	50/50	*22/130	39/80
$n = 50$		4	33	4	18		5	35	4	23		11	37	5	26		12	43	22	32
	Rqr = 0.47–1.2					*Rqr = 0.35–0.84*					*Rqr = 0.25–0.60*					*Rqr = 0.16–0.41*				
$n = 10$	355	20/23	*43/48	17/18	38/40	322	21/24	45/49	18/21	39/41	279	*21/29	47/48	27/64	43/45					
$n = 25$	377	*11/18	*32/39	10/12	25/28	366	*12/19	*34/40	11/14	26/31	298	16/20	38/40	15/20	35/35	268	24/29	42/41	82/100	44/43
$n = 50$		7	22	7	18		8	25	8	20		10	29	10	34		18	31	39	44
	Rqr = 1.2 – 3.8					*Rqr = 0.84–2.1*					*Rqr = 0.60–1.4*					*Rqr = 0.41–0.92*				
$n = 10$	301	32/35	57/52	26/30	*45/64	236	30/37	55/52	28/35	*44/61	213	31/36	53/47	63/53	46/59					
$n = 25$	304	22/27	47/44	19/18	31/42	243	*21/27	44/46	20/17	*29/43	246	23/25	46/45	27/27	*33/46	187	*29/45	45/41	150/170	43/98
$n = 50$		16	37	13	23		15	36	14	21		15	36	17	23		22	40	94	38
	Rqr > 3.8					*Rqr > 2.1*					*Rqr = 1.4–3.7*					*Rqr > 0.92*				
$n = 10$	134	55/75	65/61	*37/65	*77/130	172	62/67	73/63	46/62	*81/130	144	51/37	65/60	*130/48	*70/120					
$n = 25$	132	34/44	53/43	*25/42	*40/94	148	35/42	*57/44	27/32	*41/96	143	31/38	54/50	*77/38	38/58	164	44/40	56/57	*240/68	*37/60
$n = 50$		25	49	18	26		26	51	19	28		25	50	46	28		32	48	200	28
											Rqr > 3.7									
$n = 10$											103	93/80	93/77	240/58	110/99					
$n = 25$											54	50/53	53/45	*200/36	*57/140					
$n = 50$												33	43	170	39					

The number before each slash is the percentage rmse from the simulation results, and the number after each slash is the percentage rmse from the verification results. There are no verification results for sample size $n = 50$. The value of N is the number of verification data sets falling in each class and from which the rmse was calculated. In the simulation, data were censored at percentiles of the parent distribution, and in the verification, data sets were censored at the sample percentiles.

*Significantly different at $\alpha = 0.05$.

TABLE 4. Rank Correlations, r, Between Simulation rmses and Verification rmses for Each *rqr* Class

Degree of Censoring (As Percentage)	Parameter			
	Mean	Standard Deviation	Median	Interquartile Range
20	0.90* ($n = 8$)	0.80* ($n = 8$)		
40	0.90* ($n = 8$)	0.72* ($n = 8$)		0.87* ($n = 8$)
60	0.95* ($n = 10$)	0.58 ($n = 10$)	0.55 ($n = 10$)	0.75* ($n = 10$)
80	0.60 ($n = 4$)	0.95* ($n = 4$)	0.40 ($n = 4$)	0.40 ($n = 4$)
All censoring levels combined	0.89* ($n = 30$)	0.71* ($n = 30$)	0.37 ($n = 14$)	0.63* ($n = 22$)

*Here r is significantly different from 0.00 at $\alpha = 0.05$.

had been LR, based on its having the smallest sum of rmse ranks over all four distributional parameters, four censoring levels, and three sample sizes. By the same criteria, LR and UN tied for the best method using verification data. Also applying the same criteria, but separately for the moment parameters (mean and standard deviation) and the percentile parameters (median and interquartile range), LR produced the lowest summed rmse rank for the moment parameters and LM for the percentile parameters for both the simulated and verification data.

Verification sets were then classified by relative quartile range (*rqr*), the interquartile range of uncensored observations divided by the detection limit, and rmses were calculated for each *rqr* class. Ranks of method rmses were again separately summed for the moment and percentile parameters over both $n = 10$ and $n = 25$ sample sizes. Although for individual *rqr* classes within a censoring level a method other than LR or LM might have a smaller rmse, no rmses were significantly (t test at $\alpha = 0.05$) lower than those of LR for the moment parameters and of LM for the percentile parameters. Therefore for every *rqr* class these two methods are either best, or not significantly different from the best, and no significant reduction in error would result from selecting separate methods for each *rqr* class. This method selection exactly follows that of the simulation study.

Rmses using LR and LM are compared for this verification study with those of the prior simulation study in Table 3. The magnitudes of rmses are in most cases quite similar. Of the 130 pairs, t tests showed that 62% of the errors from the simulation study were not significantly different than errors when using actual water quality data.

There are several reasons why rmses from the verification study might not match those of the simulation study. First, during verification, water quality population values were approximated by sample estimates from relatively small data sets ($n > 50$). The differences between these estimates and the true population values introduce errors of unknown direction as compared with simulation results. Second, the small subsamples of $n = 10$ or $n = 25$ represent substantial portions of the complete data sets, lowering rmses from their true value. As a result, similarity in magnitude between simulation and verification rmses may be of limited importance.

Of importance is the similarity in effect of *rqr* classification on rmses of simulation and verification data. In order for the *rqr* classification to aid in estimating errors, both simulation and verification rmses should covary, positively, by *rqr* class. To measure this, rank correlation coefficients between simulation and verification rmses were calculated for each parameter. Ranks were first independently assigned for each study within each sample size and censoring level, and then combined. The results (Table 4) show that the relative magnitudes of rmses for the simulation and verification data are correlated, except for estimates of the median. Verification rmses for the median are low in relation to simulation data for the high censoring and high *rqr* classes (Table 3). This may be due to the lack of distributions in the verification data with as many extreme values as the high cv gamma distributions included in the simulation study [see *Gilliom and Helsel*, this issue].

The verification results are strong evidence that the previous simulation study led to optimal choice of estimation methods for the mean, standard deviation, median, and interquartile range of censored water quality data sets. Furthermore, the verification results show that the *rqr* classification system developed from simulation studies provides an effective means of distinguishing between data sets originating from different types of parent distributions.

CONFIDENCE INTERVALS FOR PARAMETER ESTIMATES

The most common objective in estimating distributional parameters from censored data is to reproduce the parameters of the parent distribution; for example, the true mean or median concentration for a particular river and time period. When making estimates of population parameters from censored data, evaluation of the reliability or confidence intervals of such estimates is an important step.

Estimation of confidence intervals requires estimates of rmse and bias. We believe that the simulation results yield more appropriate estimates of rmse and bias than do the verification results. As was previously discussed, the verification results are based on imperfect estimates of population values, and the small subsamples that were censored represent a substantial portion of the larger sample used to estimate population values. Moreover, the simulation studies included a wide range of distribution types chosen to be similar in shape to distributions of trace constituent concentrations in water [*Gilliom and Helsel*, this issue]. The verification data sets may not have represented as wide a range in distribution shapes, as only uncensored data sets were, of necessity, chosen.

The method described below for estimating confidence intervals of population parameters estimated from censored data requires three assumptions as follows.

1. The censored data are from a population that is equally likely to be similar in shape to any 1 of 16 parent distributions used in the simulation study.
2. The percentage of a data set that is censored equals the population percentile associated with the value of the detection limit.
3. Relative errors in estimated population parameters, the error of an estimate divided by the true value, can be approximated by a log-normal distribution.

The first assumption is that the 16 parent distributions used in the simulation studies appropriately represent the range and proportional contributions of different types of distributions of actual trace level water quality data. We feel confident that the range of possible shapes were included [*Gilliom and Helsel*, this issue, Figure 1]. Though there is unresolvable uncertainty about the true proportional representation of each type of distribution, the *rqr* classification system reduces this potential effect on error estimates by grouping data from similar distributions.

The assumption that the percentage of a data set that is censored equals the population percentile of the detection limit is required in order to select the proper rmse and bias

TABLE 5. Percentages of Parameter Estimates From Small Censored Samples of Simulated Data That Fell Within Computed Confidence Intervals

Method	Censored at 20th Percentile				Censored at 40th Percentile				Censored at 60th Percentile				Censored at 80th Percentile			
	LR		LM		LR		LM		LR		LM		LR		LM	
	$\bar{x}$	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*			
$n = 10$	95	89	95	91	95	87	92	92	94	81	98	91				
$n = 25$	96	87	81	93	95	89	95	95	93	89	94	94	95	85	95	92
$n = 50$	95	82	98	94	94	85	95	96	95	85	94	95	94	87	99	94
	Rqr = 0.47–1.2				*Rqr = 0.35–0.84*				*Rqr = 0.25–0.60*				*Rqr = 0.16–0.41*			
$n = 10$	93	88	93	92	93	86	94	92	94	86	99	93				
$n = 25$	93	94	93	94	94	92	94	95	95	92	94	95	95	91	99	95
$n = 50$	95	94	94	95	95	94	95	95	94	94	95	95	94	93	99	96
	Rqr = 1.2–3.8				*Rqr = 0.84–2.1*				*Rqr = 0.60–1.4*				*Rqr = 0.41–0.92*			
$n = 10$	93	88	92	93	93	88	94	94	95	89	98	95				
$n = 25$	91	93	94	93	94	90	95	93	95	89	94	96	94	91	96	95
$n = 50$	94	93	93	95	94	93	95	94	93	91	94	95	93	92	97	95
	Rqr > 3.8				*Rqr > 2.1*				*Rqr = 1.4–3.7*				*Rqr > 0.92*			
$n = 10$	98	96	93	97	98	98	95	98	97	98	98	96				
$n = 25$	96	95	94	96	96	97	93	95	95	97	96	95	96	98	98	94
$n = 50$	95	96	95	95	95	96	94	95	95	97	96	94	95	97	98	94
									Rqr > 3.7							
$n = 10$									98	97	99	99				
$n = 25$									97	97	98	92				
$n = 50$									96	97	97	88				

$\alpha = 0.05$; therefore percentages should equal 95.

values from simulation results. The simulation results are organized according to the population percentile representing the detection limit. The percentage of a data set that is censored is a sample estimate of the percentile of the detection limit, and its reliability is dependent primarily on sample size.

The assumption that relative errors are lognormally distributed was made because some probability distribution of errors must be specified to construct confidence intervals tighter than those given by Chebyshev's inequality. Box plots of errors suggested a lognormal distribution and such a distribution appeared reasonable because the fractional error,

$$e = (\bar{x} - \mu)/\mu$$

has a lower bound of -1.0, while having no upper bound. The validity of the assumed lognormal distribution of errors was directly tested by a simulation experiment. Five hundred data sets of each sample size, $n = 10$, $n = 25$, and $n = 50$, were generated from each of the 16 parent distributions described by *Gilliom and Helsel* [this issue] and censored at the 20, 40, 60, and 80th population percentiles. Sample estimates for each data set at each censoring level were made using the LR method for $\bar{x}$ and s and LM for m and iqr. Confidence intervals for each estimate at $\alpha = 0.05$ were computed as described below, and the actual frequencies with which the true population values fell within those intervals were evaluated. Results in Table 5, based on 1,000–2,000 data sets for most combinations of censoring level, sample size, and rqr class, show that the assumed lognormal distribution is generally a good approximation of the error distribution. Only for the standard deviation for the lowest three rqr classes at each censoring level and sample size is there a consistent tendency to underestimate the width of the confidence intervals.

Under the above assumptions, confidence intervals for estimates of the mean, standard deviation, median, and interquartile range may all be similarly estimated. Derivation of equations for confidence intervals are given below, using the mean as an example.

The fractional errors for estimates of the mean

$$e_i = \frac{\bar{x}_i - \mu}{\mu} \tag{2}$$

where $\bar{x}_i$ is the estimate of the mean for the ith data set and μ is the true population mean, are assumed to be lognormally distributed with mean μ_e, variance ${\sigma_e}^2$, and lower limit of -1.0. The expected value $\mu_e = E[e_i] = b$, where b is the fractional bias of estimates from the censored samples of the simulation study, is

$$b = \sum_{i=1}^{N} \frac{\bar{x}_i - \mu}{\mu} \Big/ N \tag{3}$$

and N is the number of data sets used in the simulation. The variance ${\sigma_e}^2$ is calculated as

$${\sigma_e}^2 = \text{rmse}^2 - b^2 \tag{4}$$

where rmse values are again those from the simulation.

$$\text{rmse}^2 = \sum_{i=1}^{N} \left(\frac{\bar{x}_i - \mu}{\mu} \right)^2 \Big/ N \tag{5}$$

The values of $y_i = \ln (e_i + 1.0)$ are normally distributed, with

$${\sigma_y}^2 = \ln \left(1.0 + \frac{\sigma^2 e}{(b + 1)^2} \right) \tag{6}$$

and

$$\mu_y = \ln (b + 1.0) - 0.5{\sigma_y}^2 \tag{7}$$

A $(1 - \alpha)$ confidence interval for μ is therefore given by

$$\bar{x} \exp [-\mu_y - z_{\alpha/2}\sigma_y] \leq \mu \leq \bar{x} \exp [-\mu_y + z_{\alpha/2}\sigma_y] \tag{8}$$

where z is the standard normal variate.

TABLE 6. Bias of Best Estimation Methods as Percentage of True Value

Method	Censored at 20th Percentile				Censored at 40th Percentile				Censored at 60th Percentile				Censored at 80th Percentile			
	LR		LM		LR		LM		LR		LM		LR		LM	
	$\bar{x}$	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*		
$n = 10$	−3	−34	−3	−13	−2	−39	−2	−13	4	−46	5	−15				
$n = 25$	0	−21	0	−10	−1	−22	−1	−4	−1	−28	−1	−5	9	−39	8	−14
$n = 50$	0	−17	0	−1	0	−17	−1	4	0	−23	−1	1	3	−32	2	−9
		Rqr = 0.47–1.2				*Rqr = 0.35–0.84*				*Rqr = 0.25–0.60*				*Rqr = 0.16–0.41*		
$n = 10$	−5	−23	−5	−6	−5	−23	−2	−5	1	−23	10	0				
$n = 25$	−1	−11	−2	−7	−1	−10	−1	−3	−3	−10	−1	1	3	−12	18	5
$n = 50$	0	−7	−2	−1	−1	−6	−1	1	−3	−3	−2	9	−3	−5	3	8
		Rqr = 1.2–3.8				*Rqr = 0.84–2.1*				*Rqr = 0.60–1.4*				*Rqr = 0.41–0.92*		
$n = 10$	−8	−29	−7	−4	−4	−29	0	−1	4	−25	24	7				
$n = 25$	−4	−18	−4	−9	−4	−20	−2	−9	−3	−17	2	−4	1	−9	43	8
$n = 50$	−2	−12	−3	−3	−2	−13	−2	−4	−2	−12	1	−2	−3	−7	19	5
		Rqr > 3.8				*Rqr > 2.1*				*Rqr = 1.4–3.7*				*Rqr > 0.92*		
$n = 10$	15	−16	2	35	23	−6	10	37	17	−13	56	30				
$n = 25$	7	−16	0	10	9	−12	1	10	4	−12	19	2	15	8	90	7
$n = 50$	5	−10	−3	6	7	−9	−1	7	4	−9	7	2	7	0	78	−2
										Rqr > 3.7						
$n = 10$									58	31	130	51				
$n = 25$									25	7	130	4				
$n = 50$									12	2	130	7				

To calculate confidence intervals, μ_y and σ_y are obtained from (6) and (7). The bias, b, from the simulation study is reported in Table 6. The error variance σ_e^2 is calculated in (4) using both bias from Table 6 and rmse from the simulation results reported in Table 3. The smaller rmses following classification by *rqr*, as found in the simulation study, allow shorter confidence intervals on parameter estimates than would be possible without *rqr* classification. Equation (8) can be used for any of the four distributional parameters estimated, with μ and x replaced by the population parameter and sample estimate for the standard deviation, median, or interquartile range.

The above procedure is illustrated by example in the appendix, where a 95% confidence interval for the mean is calculated. Note that neither the sample size nor the percentage of the data censored exactly correspond to conditions represented in Tables 3 and 6; this will usually be the case. One can use the rmse and bias values for the closest censoring and

TABLE 7. Rmses of Best Estimation Methods When Classified by *rqr* as Percentage of Uncensored Sample Estimate

Method	20% Censoring				40% Censoring				60% Censoring				80% Censoring			
	LR		LM		LR		LM		LR		LM		LR		LM	
	$\bar{x}$	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*		
$n = 10$	4	25			7	31		38	14	44	10	46	*	*	*	*
$n = 25$	4	27			5	29		28	8	35	5	36	21	39	66	50
$n = 50$	3	28			4	31		21	6	35	3	27	11	40	14	34
		Rqr = 0.47–1.2				*Rqr = 0.35–0.84*				*Rqr = 0.25–0.60*				*Rqr = 0.16–0.41*		
$n = 10$	3	10			7	14		22	15	27	16	43	*	*	*	
$n = 25$	2	7			5	10		17	10	18	12	42	24	25	130	61
$n = 50$	2	6			4	8		13	8	17	7	34	16	23	28	50
		Rqr = 1.2–3.8				*Rqr = 0.84–2.1*				*Rqr = 0.60–1.4*				*Rqr = 0.41–0.92*		
$n = 10$	3	5			6	8		18	14	14	54	43	*	*	*	*
$n = 25$	2	3			4	5		11	10	9	26	27	23	19	180	66
$n = 50$	1	2			3	4		8	7	6	13	16	19	17	73	47
		Rqr > 3.8				*Rqr > 2.1*				*Rqr = 1.4–3.7*				*Rqr > 0.92*		
$n = 10$	2	2			4	3		12	11	7	380	30	*	*	*	*
$n = 25$	1	1			3	2		7	7	4	44	14	18	6	170	38
$n = 50$	1	1			2	2		5	6	3	27	10	16	5	150	20
										Rqr > 3.7						
$n = 10$									7	3	290	27				
$n = 25$									5	2	92	11				
$n = 50$									5	2	64	7				

*Only 2 samples remain after censoring.

TABLE 8. Bias When Estimating Sample Statistics as Percentage of the Uncensored Sample Estimate

Method	20% Censoring				40% Censoring				60% Censoring				80% Censoring			
	LR		LM		LR		LM		LR		LM		LR		LM	
	$\bar{x}$	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*			
$n = 10$	1	−9			2	−11		4	7	−23	4	−16	*	*	*	*
$n = 25$	1	−11			0	−9		9	2	−14	1	3	7	−18	17	−6
$n = 50$	1	−12			0	−10		6	1	−16	0	1	5	−27	5	−13
	Rqr = 0.47–1.2				*Rqr = 0.35–0.84*				*Rqr = 0.25–0.60*				*Rqr = 0.16–0.41*			
$n = 10$	1	−3			2	−4		0	3	0	5	6	*	*	*	*
$n = 25$	1	−3			1	−2		3	0	1	2	9	3	2	28	14
$n = 50$	1	−3			0	−2		1	−2	3	0	10	−1	5	6	12
	Rqr = 1.2 – 3.8				*Rqr = 0.84–2.1*				*Rqr = 0.60–1.4*				*Rqr = 0.41–0.92*			
$n = 10$	1	−2			1	−2		0	1	0	11	2	*	*	*	*
$n = 25$	1	−1			1	−1		1	1	−1	3	8	−2	5	40	21
$n = 50$	0	−1			1	−1		−1	1	−1	3	−1	−1	3	17	10
	Rqr > 3.8				*Rqr > 2.1*				*Rqr = 1.4–3.7*				*Rqr > 0.92*			
$n = 10$	0	−1			0	0		2	−3	2	21	6	*	*	*	*
$n = 25$	0	−1			0	0		2	0	0	6	5	−6	2	44	16
$n = 50$	0	0			0	−1		0	0	0	5	0	−1	0	41	3
									Rqr > 3.7							
$n = 10$									−3	1	37	9				
$n = 25$									1	0	58	1				
$n = 50$									3	−1	32	−3				

*Only 2 samples remain after censoring.

sample size represented in the table, interpolate, or choose conservatively high rmses by using values for the next highest censoring level and next smallest sample size.

SAMPLE STATISTICS: ESTIMATION AND CONFIDENCE INTERVALS

For some applications, estimates of sample statistics rather than population parameters might be desired from censored data. Uncensored water quality data are summarized by their sample statistics, and comparisons between these data and censored data should be on an equal basis.

New Simulation Study

To determine how well the eight methods evaluated by *Gilliom and Helsel* [this issue] estimate sample statistics, a new simulation study was performed. Distributional shapes and other criteria are identical to the previous simulation study. However, rmses and bias were calculated (using the mean for example) as:

$$\text{rmse} = \left[\sum_{i=1}^{N} \left(\frac{\bar{x}_i - \bar{x}_0}{\bar{x}_0} \right)^2 \Big/ N \right]^{0.5}$$

$$\text{bias} = \sum_{i=1}^{N} \left(\frac{\bar{x}_i - \bar{x}_0}{\bar{x}_0} \right) \Big/ N$$

where $\bar{x}_0$ is the sample mean for the uncensored data set (replacing μ), and the other parameters are as previously given. Censoring was at the 20, 40, 60, and 80th percentiles of each simulated sample (type II censoring), as opposed to percentiles of the parent population in the first simulation study (type I censoring). This was to facilitate comparison with the verification results.

Best methods for the moment and percentile parameters in this new simulation study were LR and LM, respectively, based on the sum of method rankings over all censoring levels. The overall best method was LR. Best performing methods for estimating sample statistics were thus identical to those for estimating population parameters. However, the magnitudes of rmses differ from those for population parameters. Rmses of sample estimates in Table 7 can be compared to those of the $n = 10$ and $n = 25$ population parameters presented above the slashes in Table 3. Rmses are generally smaller when estimating sample statistics. Therefore confidence intervals around the LR or LM estimate are smaller for inclusion of the uncensored sample statistic as compared to the population parameter. Rmses for the moment sample statistics decrease with increasing *rqr* class, the opposite trend from that of the population parameters. This is due to the greater influence of the higher observations on the sample mean and standard deviation. These higher observations remain after censoring, producing a more accurately estimated sample statistic while indicating much less about the population parameter. Confidence intervals for sample statistics can be computed using the same relationships given for population parameters, but using the rmses in Table 7 and the bias results in Table 8.

Verification of Sample Statistic Estimates

To verify the new simulation results, uncensored trace metal data sets from the NASQAN network were censored (type II) at the 20, 40, 60, and 80th sample percentiles and errors were calculated by comparison to the uncensored sample estimates. Table 9 lists the water quality parameters chosen and the number of data sets for each. Sample sizes ranged from 10 to 40 observations. Eleven other trace constituents had no data sets which contained only uncensored observations and were not used. In order to obtain a larger number of data sets, iron and manganese data were included even though they are not usually found at "trace" levels.

Trace metal data sets containing 10–20 observations were combined into one group, representing sample sizes generally comparable to $n = 10$ simulation results. Data sets having fewer than three data points after censoring were deleted. A second group of data sets having from 21 to 40 observations

TABLE 9. Trace Constituents Used to Estimate Sample Statistics

Parameter	Number of Data Sets	
	$n = 10$–20	$n = 21$–40
Arsenic	7	100
Dissolved arsenic	3	63
Barium	5	0
Boron	11	3
Dissolved boron	19	7
Copper	1	13
Dissolved copper	1	5
Lead	0	17
Nickel	9	3
Zinc	1	32
Dissolved zinc	0	2
Iron	12	273
Dissolved iron	4	68
Manganese	11	180
Dissolved manganese	0	15

was formed for comparison to $n = 25$ simulation results. The eight estimation methods were applied to this data. Again, LR proved the best overall method. LR was best for the moment parameters and LM was best for the percentile parameters, based on the rank criteria given previously.

Rmses are presented by *rqr* class in Table 10. Comparison of tables 7 and 10 indicate again that simulation results produced rmses similar to those for actual trace water quality data. Only median estimates for 60 and 80% censoring appear different, with simulation rmses higher than actual. This is perhaps due to the inclusion of larger sample sizes in the actual trace-data estimates, with the simulation results representing conservative error estimates based only on $n = 10$ or $n = 25$.

SUMMARY AND CONCLUSIONS

The eight methods for estimating population parameters from censored data sets evaluated by *Gilliom and Helsel* [this issue] were applied to uncensored suspended sediment and nutrient data having large sample sizes ($n > 50$). Selection of the estimation method that was best overall, best for moment and percentile parameters separately, and best within every *rqr* class exactly follows those of the simulation study. The log regression method (LR) produced lowest rmses for the moment parameters. Rmses are similarly affected by *rqr* classification in both studies for the mean, standard deviation, and interquartile range, verifying the effectiveness of *rqr* in separating distributions which produce like errors. The differences in rmses for the median are attributed to the presence of high cv gamma distributions in the simulation study whose equivalent in the verification data, if originally present, may have contained censored values and would have therefore been excluded.

Confidence intervals for parameter estimates can be estimated using rmse and bias results from the simulation study. Fractional errors (estimate error divided by the true value) are assumed to be lognormally distributed. Simulation experiments showed that the assumption is a good approximation for all parameters except the standard deviation for data sets with low *rqr* values, for which the widths of confidence intervals were slightly underestimated. The increased accuracy of rmse estimates after *rqr* classification allow shorter confidence intervals to be constructed than would be possible without classification.

Errors in estimating statistics of uncensored samples rather than population parameters were also evaluated. Best methods for estimating sample statistics were LR and LM, respectively, for the moment and percentile parameters. Rmses were almost always smaller when estimating sample statistics than for population parameters (LM median estimates occasionally have greater rmses), and were sometimes much smaller. Therefore estimates of uncensored sample statistics are identical to those of population parameters, but have shorter confidence intervals.

The results of the present study and the companion study by *Gilliom and Helsel* [this issue] form the basis for making the best possible estimates of either population parameters or sample statistics from censored water-quality data. Moreover, they provide the means for making quantitative assessments of the reliability of those estimates, expressed as confidence bounds. The LR method for moment parameters and LM method for percentile parameters should be the methods of choice when estimating distributional parameters for censored trace level water quality data.

TABLE 10. Rmses of Best Estimation Methods for Trace Data in Percent of Uncensored Sample Estimate

Method	20% Censoring				40% Censoring				60% Censoring				80% Censoring			
	LR		LM		LR		LM		LR		LM		LR		LM	
	$\bar{x}$	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*			
$n = 10$–20	4	23			9	31		57	15	39	10	73	30	53	32	55
$n = 21$–40	2	10			5	15		44	12	24	16	43	23	28	73	45
	Rqr = 0.47–1.2				*Rqr = 0.35–0.84*				*Rqr = 0.25–0.60*				*Rqr = 0.16–0.41*			
$n = 10$–20	3	8			7	13		23	11	15	14	29	15	16	40	37
$n = 21$–40	2	5			5	9		25	10	12	12	45	21	17	63	51
	Rqr = 1.2 -- 3.8				*Rqr = 0.84–2.1*				*Rqr = 0.60–1.4*				*Rqr = 0.41 – 0.92*			
$n = 10$–20	3	6			6	8		11	12	12	16	16	20	20	34	51
$n = 21$–40	2	3			5	5		17	10	8	17	30	22	13	47	41
	Rqr > 3.8				*Rqr > 2.1*				*Rqr > 1.4–3.7*				*Rqr > 0.92*			
$n = 10$–20	1	1			4	3		9	7	6	11	21	28	9	86	99
$n = 21$–40	1	1			3	2		14	9	5	24	30	22	8	150	72
									Rqr > 3.7							
$n = 10$–20									6	2	28	39				
$n = 21$–40									6	2	52	25				

APPENDIX: CALCULATION OF $1 - \alpha$ CONFIDENCE INTERVAL ($\alpha = 0.05$) FOR ESTIMATE OF THE MEAN FROM CENSORED DATA

Data set: ND, ND, ND, ND, ND, 6, 6, 6, 8, 10, 10, 10, 11, 20

Detection limit = 5.0 $n = 14$

Percent censoring = 36% $rqr = \dfrac{10.5-6}{5} = 0.90$

Estimates using LR method:

$$\text{mean} = 7.42$$

$$\text{standard deviation} = 4.66$$

From Tables 3 and 6, for

$n = 10$ percent censoring = 40% LR method,

rmse = 0.30 bias = -0.04

$\sigma_e^2 = 0.088$ $\sigma_y^2 = 0.092$ $\mu_y = -0.087$

$$7.42 \exp [+0.087 - (1.96)(0.303)] \leq \mu \leq 7.42 \cdot \exp [0.087 + (1.96)(0.303)]$$

$$4.47 \leq \mu \leq 14.64$$

Acknowledgment. The authors thank Edward J. Gilroy for several helpful 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 Brown, J. A. C., *The Lognormal Distribution*, 176 pp., Cambridge University Press, New York, 1957.

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

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

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.)