WATER RESOURCES RESEARCH, VOL. 32, NO. 4, PAGES 1005–1012, APRIL 1996

Estimation of moments and quantiles using censored data

Charles N. Kroll and Jery R. Stedinger
School of Civil and Environmental Engineering, Cornell University, Ithaca, New York

Abstract. Censored data sets are often encountered in water quality investigations and streamflow analyses. A Monte Carlo analysis examined the performance of three techniques for estimating the moments and quantiles of a distribution using censored data sets. These techniques include a lognormal maximum likelihood estimator (MLE), a log-probability plot regression estimator, and a new log-partial probability-weighted moment estimator. Data sets were generated from a number of distributions commonly used to describe water quality and water quantity variables. A "robust" fill-in method, which circumvents transformation bias in the real space moments, was implemented with all three estimation techniques to obtain a complete sample for computation of the sample mean and standard deviation. Regardless of the underlying distribution, the MLE generally performed as well as or better than the other estimators, though the moment and quantile estimators using all three techniques had comparable log-space root mean square errors (rmse) for censoring at or below the 20th percentile for samples sizes of $n = 10$, the 40th percentile for $n = 25$, and the 60th percentile for $n = 50$. Comparison of the log-space rmse and real-space rmse indicated that a log-space rmse was a better overall metric of estimator precision.

Introduction

When a data set contains some observations within a restricted range of values but otherwise not measured, it is called a censored data set [*Cohen*, 1991]. Censored data sets are commonly found in the fields of water quality, where laboratory measurements of contaminant concentrations are often reported as "less than the detection limit." Censored data sets are also found in water quantity analyses when river discharges less than a measurement threshold level are reported as zero. In some regions, historical river discharge records report over half the annual minimum flows as zero [*Hammett*, 1984]. These discharges may have been zero, or they may have been between zero and the measurement threshold and thus reported as zero. Of concern is how to efficiently estimate moments, quantiles, and other descriptive statistics of the underlying continuous distribution using such censored data sets.

The situation where all data below a fixed value are censored is referred to as type I censoring. With type I censoring, the number of values censored is a random variable. With type II censoring, a fixed number of data points are always censored and the censoring threshold is a random variable [*David*, 1981]. Censored water quality and water quantity data should resemble type I censoring because the censoring threshold is fixed by the measurement technology and the physical setting.

A number of studies have suggested the use of simple "replacement" techniques for estimating the mean and standard deviation of type I censored data sets [*Cohen and Ryan*, 1989; *Newman et al.*, 1990]. These techniques replace all the censored observations with some value between zero and the detection limit. *Gilliom and Helsel* [1986] and *Helsel and Gilliom* [1986] examined the performance of a variety of techniques to estimate the mean, standard deviation, median, and interquartile range using type I censored water quality data. They showed that more sophisticated statistical techniques performed better than these simple "replacement" methods. In particular, the log-probability plot regression method provided the best estimators of the mean and standard deviation, while the lognormal maximum likelihood method provided the best estimators of the median and interquartile range. Estimation of quantiles other than the median was not considered by Gilliom and Helsel. *Helsel and Cohn* [1988] extended Gilliom and Helsel's work to data sets with several censoring thresholds.

This study extends the work of Gilliom and Helsel to the estimation of several quantiles and considers new estimators. The log-probability plot regression method (LPPR) and the lognormal maximum likelihood method (MLE) are evaluated along with a new method based on partial probability-weighted moments (PPWM). As with the MLE and LPPR estimators, our PPWM estimator assumes that the data are described by a lognormal distribution. It employs with the logarithms of the flow data the censored-sample probability-weighted moment (PWM) estimators derived by *Wang* [1990] to obtain the parameters of a lognormal distribution. Wang employed his estimators in real space to fit a generalized extreme value (GEV) distribution. The performance of probability weighted moment estimators with complete samples has been examined for a number of distributions, and in many cases, PWM estimators of the higher moments and quantiles of a distribution have performed favorably with product-moment and maximum likelihood estimators [*Landwehr et al.*, 1979; *Hosking et al.*, 1985; *Hosking and Wallis*, 1987]. PWM estimators are linear combinations of the observations and thus are less sensitive to the largest observations in a sample than product-moment estimators that square and cube the observations. The merit of probability-weighted moment estimators with censored samples has yet to be analyzed.

This study focuses on estimation of the mean, standard deviation, and interquartile range of a distribution, as well as quantiles with nonexceedance probabilities of 10% and 90%. A "robust" fill-in method is implemented with each estimation

Paper number 95WR03294.
0043-1397/96/95WR-03294$05.00

technique to obtain a complete sample for computation of the sample mean and variance. *Gilliom and Helsel* [1986] used this "robust" fill-in method only with a log-probability plot regression estimator. In this study this method is also used with the lognormal maximum likelihood and partial probability-weighted moment estimators. Two different metrics are used to compare estimators. Data are generated from distributions commonly observed in the water quality and water quantity fields, including three distributions not considered by Gilliom and Helsel; their extreme case for the gamma distribution (coefficient of variation = 2.0) was omitted.

Estimation Techniques

All three estimation techniques make the assumption that the underlying distribution of the data is lognormal. *Helsel and Hirsch* [1992, p. 360] observe that the lognormal distribution has a flexible shape, and they provide a reasonable description of many positive random variables with positively skewed distributions. The lognormal distribution has been shown to be a good descriptor of low river flows [*Vogel and Kroll*, 1989] and water quality data [*Gilliom and Helsel*, 1986].

Lognormal Maximum Likelihood Estimator

Consider an ordered censored data set $X_1 \leq X_2 \cdots \leq X_c \leq X_{c+1} \cdots \leq X_n$, where the first c observations are censored and reported only as below some fixed measurement threshold. Let $Y_i = \ln(X_i)$ and let T be the log of the measurement threshold. Assuming that X is lognormally distributed and independent, the likelihood function for the data is

$$L = \frac{n!}{c!(n-c)!}\left[\Phi\left(\frac{T-\mu_Y}{\sigma_Y}\right)\right]^c \prod_{i=c+1}^{n} \frac{1}{\sigma_Y}\phi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) \quad (1)$$

where Φ and ϕ are the distribution and density function of a standard normal variate, μ_Y is the mean of the log-transformed data, and σ_Y is the standard deviation of the log-transformed data. By taking the logarithm of (1) and setting the partial derivatives with respect to μ_Y and σ_Y to zero, one can solve for the maximum likelihood estimators (MLE) $\hat{\mu}_Y$ and $\hat{\sigma}_Y$ [*Cohen*, 1991].

Log-Probability Plot Regression Method

Again consider a log-transformed censored data set where the first c of the n data values are censored. Plotting positions for the uncensored observations are

$$p_i = \frac{c}{n} + \left(\frac{n-c}{n}\right)\left(\frac{i-\frac{3}{8}-c}{n+\frac{1}{4}-c}\right) \qquad i = c+1, \cdots, n \quad (2)$$

where i is the rank of the ith flow. This is the Blom-based plotting position for censored data developed by *Hirsch and Stedinger* [1987]. *Liu and Stedinger* [1991] found that quantile estimators with the Hirsch-Stedinger Blom-based censored data plotting position (equation (2)) had a smaller root mean square error than estimators with a standard complete sample Weibull plotting position $[i/(n+1)]$ when censored data were present. *Gilliom and Helsel* [1986] used the standard complete sample Weibull plotting position in their LPPR estimator, while *Helsel and Cohn* [1988] used a Weibull-based plotting position with the Hirsch-Stedinger censored data plotting position.

For the data above the threshold the logarithm of the ordered values, Y_i, are regressed against the corresponding "normal scores" corresponding to the model

$$Y_i = \hat{\mu}_Y + \hat{\sigma}_Y\Phi^{-1}(p_i) + \varepsilon_i \qquad i = c+1, \cdots, n \quad (3)$$

where $\Phi^{-1}(p_i)$ is the inverse cumulative normal distribution function evaluated at p_i, and $\hat{\mu}_Y$ and $\hat{\sigma}_Y$ are the resulting estimators of the mean and standard deviation of the log-transformed data obtained using ordinary least squares regression. These LPPR estimators are similar to those derived by *Gupta* [1952] and have been implemented in a number of studies of estimation with censored data sets [*Gilliom and Helsel*, 1986; *Helsel and Gilliom*, 1986; *Helsel and Cohn*, 1988; *Helsel*, 1990].

Partial Probability Weighted Moments

For a variable Y, probability-weighted moments are defined as

$$\beta_r = E\{Y[F(Y)]^r\} \quad (4)$$

where $F(Y)$ is the cumulative distribution function (CDF) for Y. For a continuous random variable, PWMs can be written

$$\beta_r = \int_0^1 Y(F)\,F^r\,dF \quad (5)$$

where $F = F(Y)$ and $Y(F)$ is the inverse CDF of Y evaluated at the probability F. For a censored sample, *Wang* [1990] defined a PPWM as

$$\beta_r(F_T) = \int_{P_T}^1 Y(F)\,F^r\,dF \quad (6)$$

where $P_T = F(T)$, the probability of censoring, and T is the censoring threshold.

Assuming the data X are lognormally distributed, and $Y = \log(X)$, then Y is normally distributed. T is the log of the censoring threshold. For the normal distribution the inverse CDF for a random variable Y is

$$Y(F) = \mu_Y + \sigma_Y\Phi^{-1}(F) \quad (7)$$

An approximation to $\Phi^{-1}(F)$ is

$$\Phi^{-1}(F) \doteq 5.05[F^{0.135} - (1-F)^{0.135}] \quad (8)$$

This is a good approximation of the normal inverse CDF for $0.005 \leq F \leq 0.995$ [*Joiner and Rosenblatt*, 1971]. Substituting (8) and (7) into (6) yields

$$B_r(P_T) = \int_{P_T}^1 (\mu_Y + \sigma_Y(5.05)[F^{0.135} - (1-F)^{0.135}])\,F^r\,dF \quad (9)$$

$B_r(P_T)$ is an approximation of $\beta_r(P_T)$ based on (8). For $r = 0$,

$$B_0(P_T) = \mu_Y[g_1(P_T)] + \sigma_Y[g_2(P_T)]$$

$$g_1(P_T) = (1-P_T) \quad (10)$$

$$g_2(P_T) = \left[\frac{5.05}{1.1335}\right][1 - P_T^{1.135} - (1-P_T)^{1.135}]$$

and for $r = 1$,

$$B_1(P_T) = \mu_Y[g_3(P_T)] + \sigma_Y[g_4(P_T)]$$

$$g_3(P_T) = \left[\frac{1}{2} - \frac{P_T^2}{2}\right] \tag{11}$$

$$g_4(P_T) = 5.05\left[\frac{1}{2.135} - \frac{P_T^{2.135}}{2.135} - \frac{P_T(1-P_T)^{1.135}}{1.135} - \frac{(1-P_T)^{2.135}}{(1.135)(2.135)}\right]$$

For $X_1 \le X_2 \le \cdots \le X_n$, an unbiased estimator of $\beta_r(P_T)$ is [*Wang*, 1990]

$$b_r(P_T) = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-r)}{(n-1)(n-2)\cdots(n-r)} Y_i \tag{12}$$

where

$$\begin{aligned} Y_i &= 0 && \text{if } X_i < \exp(T) \\ Y_i &= \ln(X_i) && \text{if } X_i \ge \exp(T) \end{aligned} \tag{13}$$

Wang [1990] recommended estimating P_T as

$$\hat{P}_T = c/n \tag{14}$$

where c is the number of observation not exceeding T. Setting equations for $B_0(\hat{P}_T)$ and $B_1(\hat{P}_T)$ equal to the estimators $b_0(P_T)$ and $b_1(P_T)$, PPWM estimators of $\hat{\mu}_Y$ and $\hat{\sigma}_Y$ are obtained:

$$\hat{\sigma}_Y = \frac{b_1(P_T)g_1(\hat{P}_T) - b_0(P_T)g_3(\hat{P}_T)}{g_1(\hat{P}_T)g_4(\hat{P}_T) - g_2(\hat{P}_T)g_3(\hat{P}_T)} \tag{15}$$

$$\hat{\mu}_Y = \frac{b_0(P_T)}{g_1(\hat{P}_T)} - \frac{g_2(\hat{P}_T)}{g_1(\hat{P}_T)}\hat{\sigma}_Y \tag{16}$$

Unlike other applications of PWMs, here PPWM estimators are applied to the log-transformed data. Applying a log transformation to the data before calculating sample moments is another approach to reducing the influence of the largest observations [*Stedinger et al.*, 1993, p. 18-5]. *Hosking* [1989] developed a real-space PWM estimator for the parameters of a lognormal distribution. The simplifications that Hosking employed in his derivation cannot be adapted to real-space PPWM estimators with the lognormal distribution because the integration of (6) is over a limited domain and not from 0 to 1 as in Hosking's formulation. A real-space PPWM estimator with the lognormal distribution would require complicated integration procedures and were not explored in this study.

Estimation of Statistics

The focus of this study is the estimation of the mean (μ), standard deviation (σ), interquartile range (IQR), and quantiles with nonexceedance probabilities of 10% (X_{10}) and 90% (X_{90}). The MLE, LPPR, and PPWM estimators describe the mean and variance of the log-transformed data. The corresponding estimator of the various quantiles is

$$\hat{X}_P = \exp(\hat{\mu}_Y + z_P\hat{\sigma}_Y) \tag{17}$$

where $\hat{X}_p$ is an estimate of X_p, a quantile with an nonexceedance probability of p percent, and z_p is the inverse of the standard normal cumulative distribution function evaluated at pth percentile.

To obtain estimators of the mean and variance in real space, one could transform the log-space mean and variance into the real-space moments [*Aitchison and Brown*, 1957]. The real-space estimators would be biased due to this transformation, even if the log-space estimators are unbiased [*Finney*, 1941]. The real-space sample estimates of the mean and standard deviation are most sensitive to the largest observations, and lack of fit to these observations can produce substantial error in these estimators. Several studies have examined techniques which try to correct for this bias [*Helsel and Cohn*, 1988; *Cohn et al.*, 1989; *Newman et al.*, 1990]. *Helsel and Hirsch* [1992, pp. 360–361] note that compensating for this bias requires an assumption about distributional shape, which is impossible when the underlying distribution of the data is unknown.

A solution to this problem is to combine the observed data above the censoring threshold with estimators of the smallest observations which were censored because they fell below the measurement threshold. Estimates of the mean and standard deviation are then obtained as the sample mean and sample standard deviation of this new data set. *Helsel and Hirsch* [1992] indicated that such estimation techniques are "robust" because they perform well even when the data are not lognormally distributed.

Gilliom and Helsel [1986] only employed this "robust" method using the LPPR estimator, though this techniques could be applied with any estimator. With this technique the regression relationship is used to extrapolate the below threshold observations. The plotting position for the c censored observations are

$$p_i = \left(\frac{c}{n}\right)\left(\frac{i - \frac{3}{8}}{c + \frac{1}{4}}\right) \qquad i = 1, \cdots, c \tag{18}$$

[*Hirsch and Stedinger*, 1987]. The censored observations are then estimated by

$$X_i = \exp[\hat{\mu}_Y + \hat{\sigma}_Y\Phi^{-1}(p_i)] \qquad i = 1, \cdots, c \tag{19}$$

using p_i from (18). The mean and standard deviation of the data are estimated as the sample mean and sample standard deviation of the completed data set. This technique will be used to obtain real-space estimators of the mean and standard deviation associated with all three estimators. For each estimator, estimates of the mean and standard deviation in logarithmic space will be used in (19) to obtain estimates of the censored observations. These estimates will be combined with the uncensored observations to obtain the sample mean and sample standard deviation of the completed data set.

Data Generation

Data for this experiment was generated from a number of distributions which are commonly used to describe water quality and water quantity data. *Gilliom and Helsel* [1986] generated data from four distributions: lognormal, contaminated lognormal, gamma, and delta. The contaminated lognormal distribution they used is a combination of two lognormal distributions, X_1 which describes 80% of the distribution, and X_2 which describes 20% of the distribution. The moments of the two distributions are related by $\mu_{X_2} = 1.5\ \mu_{X_1}$, and $\sigma_{X_2}/\mu_{X_2} = 2.0\sigma_{X_1}/\mu_{X_1}$. *Gilliom and Helsel* [1986] derived the moments of this distribution. The delta distribution employed in Gilliom and Helsel was a lognormal distribution plus a point mass of 5% at zero. *Aitchison* [1955] provides a general description of

such a delta distribution. Based on uncensored water quality records, Gilliom and Helsel suggest that these four distributions adequately describe the characteristics of most water quality data. They considered each distribution with a coefficient of variation (CV) of 0.25, 0.5, 1.0, or 2.0.

While censoring could occur when recording daily or even annual maximum river flows, it is most likely to occur when recording annual minimum river flows. The true distribution of river flows is not known, but several distributions have been found to describe annual minimum flows. *Tasker* [1987] recommended the log-Pearson III and Weibull distributions for estimating at-site frequency curves in Virginia. *Condie and Nix* [1975] found a Weibull distribution provided a good fit to Canadian rivers. *Vogel and Kroll* [1989] showed that a lognormal model is reasonable for annual minimum flows in the northeastern United States.

Based on these studies, data was generated from seven distributions: lognormal, contaminated lognormal, gamma, delta, Weibull, log-Pearson III with log skew equal to 0.25, and log-Pearson III with log skew equal to 1.0. For the log-Pearson III distribution, the log skew are representative of small and large values observed with annual minimum streamflow data [*Tasker*, 1989]. The log-Pearson III is similar in shape to the contaminated lognormal distribution. Both of these distributions have a thicker upper tail than the lognormal distribution. The contaminated lognormal and delta distributions used here were the same as those employed by *Gilliom and Helsel* [1986].

For each distribution, four variants based on a CV = 0.25, 0.5, 1.0, and 2.0 were included. The mean of the flows was set to 1.0. A gamma distribution with a CV = 2.0 produces quantiles less than the median which are very close to zero (the 50th percentile of this distribution equals 0.17 and the 40th percentile of this distribution equals 0.06). Owing to the extreme character of a gamma distribution with CV = 2.0, data were not generated from this distribution. Interestingly, *Gilliom and Helsel* [1986] used results for gamma distribution with a CV = 2.0 to show that the MLE could be a poor estimator of the mean and standard deviation for nonlognormal data.

Combining all combinations of distribution and CV (except gamma with CV = 2) yields 27 different parent distributions. Five thousand samples of 10, 25, and 50 observations (n = 10, 25, and 50) were generated for each of these 27 combinations. Complete samples were generated as well as samples with censoring levels set at the 10th, 20th, 40th, 60th, and 80th percentile of the parent distribution. Data sets with less than three uncensored observations were discarded. In practice, these estimation procedures would not be performed when only one or two observations are uncensored. Data sets were generated until 5000 acceptable data sets were available.

The delta distribution used in this experiment is a lognormal distribution with a point mass at zero having a probability of 5%. Since this distribution produces data equal to zero, the estimation methods can not be performed for the case with 0% censoring, and thus that case was excluded.

Performance Measures

Estimation methods were compared using two different performance measures: the relative root mean square error in real space (R-rmse), and the root mean square error in log space (L-rmse). The bias of the estimators was also calculated, though those results are not reported. The relative root mean square error (R-rmse) of an estimator in real space was calculated as

$$\text{R-rmse} = \left[\sum_{i=1}^{N} (\hat{\theta}_i - \theta)^2 \Big/ N \right]^{1/2} \Big/ \theta \qquad (20)$$

where $\hat{\theta}_i$ is an estimate of the statistic θ and N is the number of replicates of the experiment (5000 for each parent distribution). This metric is commonly used to evaluate the performance of estimation methods and was employed by *Gilliom and Helsel* [1986] and *Helsel and Cohn* [1988].

The second criterion is the log-space rmse, defined as

$$\text{L-rmse} = \left[\sum_{i=1}^{N} \left(\ln \left(\frac{\hat{\theta}}{\theta} \right) \right)^2 \Big/ N \right]^{1/2} \qquad (21)$$

With the log-space rmse underestimation errors receives more weight than overestimation errors. This criterion was employed by *Stedinger and Cohn* [1986] and *Fill* [1994]. The real-space metric given by (20) assigns symmetric losses to over and underestimation errors of equal magnitude. The log-space metric given by (21) assigns symmetric losses to equal percentage of over and underestimation errors. It is easily shown that the two metrics are equivalent to first order for small errors.

The choice between competing estimators often requires a trade-off between bias and mean square error. In some cases, unbiased estimators may be scaled so that the resulting negatively biased estimator has a smaller mean square error than the original estimator. Consider the estimator of the variance with normal data. The traditional unbiased estimator, s^2, can be scaled by a factor $\gamma = (n - 1)/(n + 1)$, where n is the sample size, to produce a biased estimator, γs^2, with the smallest mean square error among all estimators of the form γs^2. However, if the root mean square error was divided by the expected value of the estimator, the resulting coefficient of variation would be the same for both estimators. Thus the scaled estimator has the same relative precision but is biased and hence does not represent any real improvement over the traditional estimator. The R-rmse performance criterion favors the scaled estimator over the traditional estimator, because the reduction in variance is greater than the squared increase in bias. The L-rmse criteria is not fooled by such scaling because the variance of the logarithm of an estimator is unaffected by multiplying an estimator by a fixed scalar, and any increase in the log-space bias due to scaling increases the L-rmse of an estimator.

Results

This experiment includes results for data generated from the three groups of parent distributions. The first group includes results for samples drawn from lognormal distributions (LN). Since the MLE, LPPR, and PPWM estimators all assume a lognormal distribution, of interest is the performance of these estimators when the data are generated from that distribution. The second group (water quality (WQ)) comprises data generated from distributions commonly used to describe water quality data: lognormal (LN), contaminated lognormal (CLN), gamma (g), and delta (D). These were the distributions that *Gilliom and Helsel* [1986] considered in their analysis of censored water quality data. The third group (low flow, LF) includes data generated from distributions commonly used to

Table 1. R-rmse and L-rmse of Estimators as a Percentage of the True Value for Lognormal Data

	X_{10}		X_{90}		IQR		MEAN		s.d.	
Method	R-rmse	L-rmse	R-rmse	L-rmse	R-rmse	L-rmse	R-rmse	L-rmse	R-rmse	L-rmse
					Censoring at the 0th Percentile					
MLE	24	22	23	22	24	23	23	21	42	45
LPPR	23	22	25	22	23	24	23	21	42	45
PPWM	23	22	25	22	23	24	23	21	42	45
					Censoring at the 20th Percentile					
MLE	27	26	24	23	24	24	23	21	43	46
LPPR	30	28	25	23	25	25	23	21	43	46
PPWM	28	27	25	23	25	25	23	21	43	46
					Censoring at the 60th Percentile					
MLE	47	46	24	24	29	26	23	21	45	49
LPPR	63	57	26	24	30	28	23	21	46	51
PPWM	57	60	26	24	29	28	24	22	45	49
					Censoring at the 80th Percentile					
MLE	106	72	23	20	34	29	24	20	48	53
LPPR	168	100	26	22	40	33	29	24	50	58
PPWM	72	164	29	29	45	35	32	37	49	45

Data set size is 25. Standard error of all estimates are <3% for all cases reported above.

describe annual minimum streamflows in the United States: lognormal (LN), log-Pearson III (LPIII) with log skew of 0.25 and 1.0, and Weibull (W).

For $n = 10$ and censoring at the 80th percentile, the probability of two or fewer uncensored observations is 68%. Results for censoring at the 80th percentile with $n = 10$ were therefore not considered meaningful. With $n = 10$ and censoring at the 60th percentile, the probability of two or fewer uncensored observations is 17%, while with $n = 25$ and censoring at the 80th percentile the probability is 10%. These cases were also not considered meaningful in this analysis. Numerical results for all cases are not reported here but are given by *Kroll* [1996].

Group I: Lognormal Data (LN)

Table 1 contains the relative real-space root mean square error (R-rmse) and log-space root mean square error (L-rmse) of the estimators averaged over all CV values when the underlying distribution is lognormal for $n = 25$ and censoring occurs at the 0th, 20th, 60th, and 80th percentiles. In general, the real- and log-space metrics produce the same ranking of the estimators. The exception is at high censoring, where the bias of some estimators is large. For censoring at the 80th percentile the PPWM estimator of X_{10} has a smaller real-space rmse than the MLE and LPPR estimators, but a larger log-space rmse. Note that the R-rmse and L-rmse for some estimators is smaller with censoring at the 80th percentile than the 60th percentile. This is probably due to rejecting a large number of data sets with two or fewer uncensored observations when censoring is at the 80th percentile.

Table 2. Performance of Estimators of X_{10} When CV = 1.0, Sample Size $n = 50$, and Censoring is at the 80th Percentile

Method	R-rmse	L-rmse	Mean	Median	Bias
MLE	0.35	0.36	0.28	0.27	0.04
LPPR	0.62	0.59	0.30	0.27	0.06
PPWM	0.32	1.41	0.17	0.15	−0.07

True value of $X_{10} = 0.243$.

Table 2 is presented to compare the two metrics for X_{10} estimators. When CV = 1.0, $n = 50$, and the censoring threshold at the highest level: the 80th percentile. Table 2 reports the R-rmse, L-rmse, mean, median, and bias of the estimators. The PPWM estimator has the smallest R-rmse but the largest L-rmse. The MLE estimator has the smallest L-rmse. Figure 1 illustrates the distribution of the three estimators based on the 5000 replicates. The MLE and LPPR estimators are less biased than the PPWM estimator but can also yield large overestimates. In general, a probability density function (pdf) symmetric about the true value is favorable. The distribution of the PPWM estimator resembles a scaled version of the distribution of the MLE or LPPR estimators. If one had used R-rmse as a comparison metric, the PPWM estimator would be better than the MLE and LPPR estimators for this case. Based on the pdf, mean, and median of the estimators, the MLE and LPPR estimators appear preferable to the PPWM estimator. Using L-rmse as a metric, a similar conclusion would be drawn. L-rmse gives greater weight to underestimation and less to overestimation and is not mislead by scaled estimators which may produce a reduction in R-rmse.

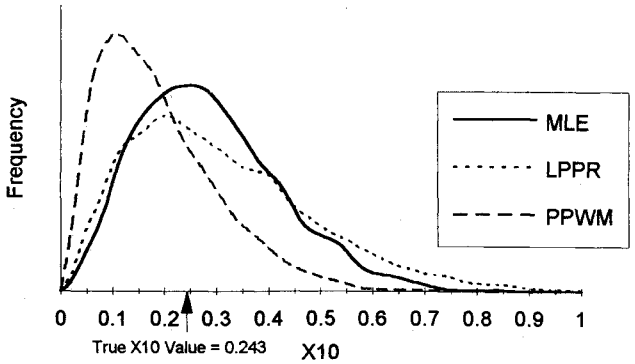

Figure 1. Probability density function of X_{10} estimators with CV = 1.0, $n = 50$, and censoring at the 80th percentile.

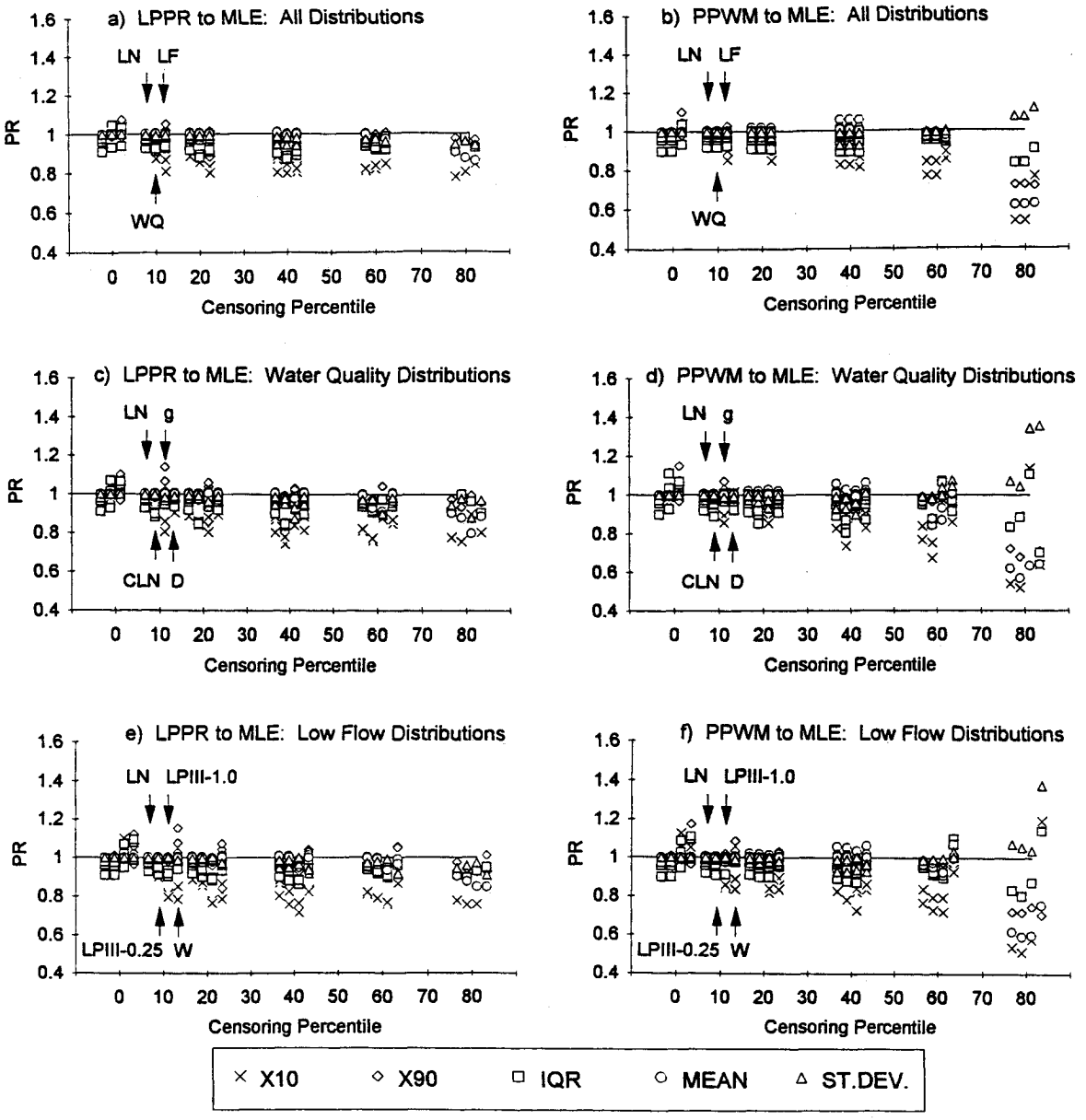

Figure 2. Performance ratios of LPPR to MLE and PPWM to MLE for lognormal, water quality, and low flow data with $n = 10$, 25, and 50.

L-rmse appears to be a better performance criterion than R-rmse for strictly positive estimators.

Based on L-rmse when the underlying distribution is lognormal over the range of cases in Table 1, the MLE is the best estimator of X_{10}, X_{90}, IQR, μ, and σ, though at extreme censoring (80th percentile) the PPWM is the best estimator of σ. All estimators had comparable L-rmse for censoring at or below the 40th percentile for $n = 25$. Although the results are not reported here, when $n = 10$, all estimators had comparable L-rmse for censoring at or below the 20th percentile, and when $n = 50$, all estimators had comparable L-rmse for censoring at or below the 60th percentile. In general, the MLE and LPPR estimators of the standard deviation have a negative bias at extreme censoring (80th percentile), while the PPWM has a positive bias. This result may be due to rejecting data sets with two or fewer uncensored observations. Since on average it overestimates as opposed to underestimates the standard deviation, the PPWM estimator of the standard deviation produces a smaller L-rmse than the L-rmse of the MLE and LPPR estimators, while the R-rmse of all estimators are almost identical.

To illustrate how well the LPPR estimator performs relative to the MLE, a performance ratio (PR) for the estimators was calculated as:

$$\text{PR} = [\text{L-rmse(MLE)}]/[\text{L-rmse(LPPR)}] \qquad (22)$$

Figure 2a is a plot of PR versus censoring percentile when the underlying distribution is lognormal. Note that the results for water quality and low flow distribution groups are also included in Figure 2. For each censoring percentile, the 15 values plotted for each group refer to the PR of the five statistics for $n = 10$, 25, and 50. To avoid bias due to rejecting a large number of samples, the cases with $n = 10$ for censoring at the 60th percentile and with $n = 10$ and 25 for censoring at the 80th percentile are omitted. As the censoring level increases,

more of the PR are less than 1, indicating a higher L-rmse for the LPPR estimator than the MLE estimator. This is especially true for estimators of X_{10}. Figure 2b is a plot of the PR for the PPWM compared to the MLE. As censoring increases, the PPWM estimators have a higher L-rmse than the MLE estimator, except for the standard deviation. The PPWM does especially poorly when estimating X_{10}. Except for estimation of the standard deviation at high censoring percentiles, the PR of the LPPR estimators are usually closer to 1 than the PR of the PPWM estimators, especially at higher censoring thresholds. This indicates that the LPPR is generally a better estimator than PPWM, though the performance differences are modest except for X_{10} estimators.

Estimation of X_{10} is interesting since the methods are generally forced to extrapolate below the censoring threshold. At low censoring, most of the information about the distribution and its lower tail is contained in the uncensored observations, rather than number of uncensored observations, and all estimation methods produce similar results. As the rate of censoring increases, less overall information about the distribution is contained in the above threshold observations and the relative amount of information provided by the uncensored observations compared to the information provided by the censoring rate decreases. These trends could be illustrated quantitatively using the Fisher information matrix [*Judge et al.*, 1985]. The MLE uses the information provided by the censored observations more efficiently than the other estimators and thus produces an estimator of X_{10} with a smaller L-rmse than the other estimators as the censoring rate increases.

With the lognormal distribution, the data in log-space are described by a normal distribution. Use of L-rmse is therefore equivalent to comparing the rmse of the estimators of X_{10} and X_{90} for the normal distribution. From this perspective, the PPWM is a real-space estimator for the normal distribution, as opposed to a log-space estimator for the lognormal distribution. The rmse of the real-space estimators for the normal distribution are the same as the L-rmse of the X_{10} and X_{90} estimators for the lognormal distribution given in Table 1. Thus for the normal distribution the rmse of the three estimators of X_{10} are generally equivalent when censoring is at or below the 10th percentile for $n = 10$, below the 20th percentile for $n = 25$, and below the 40th percentile for $n = 50$.

Group II: Water Quality Data (WQ)

Figure 2a contains the PR of the LPPR estimators to the MLE estimators for results averaged over all water quality (WQ) distributions. These results are similar to those for the lognormal distribution discussed above. Figure 2c compares the PR of LPPR estimators to the MLE estimators for each water quality distribution. Even when the underlying distribution was not lognormal, the MLE estimators generally performed better than the LPPR estimators, especially at higher censoring levels. However, the LPPR estimator of X_{90} had a smaller L-rmse than the MLE estimator when the underlying distribution was gamma and censoring was at or below the 40th percentile. The shape of a gamma distribution with a high CV value differs considerably from the shape of a lognormal distribution.

Figure 2b contains the PR of the PPWM estimators to the MLE estimators for results averaged over all water quality distributions. Figure 2d compares the PR of the PPWM estimators to the MLE estimators for the different water quality distributions. For censoring at or below the 40th percentile, the L-rmse of the PPWM estimator of the mean is slightly lower than the L-rmse of the MLE estimator. At high censoring, the PPWM is a better estimator of the standard deviation. The MLE estimator of X_{90} performs poorly at low censoring when the underlying distribution is gamma. Comparing Figures 2c and 2d, the LPPR estimators generally are as good as or better than the PPWM estimators. The exception is for estimators of the standard deviation at high censoring levels.

Group III: Low Flow Data (LF)

Figure 2a contains the PR of the LPPR estimators to the MLE estimators for results averaged over all low flow distributions. Figure 2e compares the PR of LPPR estimators to the MLE estimators for the different low flow distributions. In general, the MLE estimators have a smaller L-rmse than the LPPR estimators. The exception is the estimator of X_{90} when the underlying distribution is Weibull, a distribution whose shape differs significantly from the shape of the lognormal distribution.

Figure 2b contains the PR of the PPWM estimators to the MLE estimators for results averaged over all low flow distributions. Figure 2f compares the PR of PPWM estimators to the MLE estimators for the different low flow distributions. When censoring is high, the L-rmse of the PPWM estimator of the standard deviation is less than the L-rmse of the MLE estimator. The MLE also performs poorly when estimating X_{90} for the Weibull distribution. In general, the MLE estimators have a smaller L-rmse than the PPWM estimators. Comparing Figures 2e and 2f, the LPPR estimators generally are as good as or better the PPWM estimators, except when estimating the standard deviation at high censoring percentiles.

Conclusions

The following conclusions can be drawn from these experiments:

1. The log-space rmse (L-rmse) and the relative real-space rmse (R-rmse) produced similar ranking of the estimators, except for some estimators with large negative biases. The L-rmse places a larger penalty on underestimation errors and a smaller penalty on overestimation errors than the R-rmse. The L-rmse appears to be a better estimator performance metric than the R-rmse because it is not mislead by estimators which may represent a scaling that produces a negative bias and a smaller R-rmse but no increase in information.

2. When the estimators were tested with data drawn from a range of distributions representative of both water quality and water quantity measurements, the ranking of the estimators is generally the same. Regardless of the underlying distribution, the MLE generally performed as well as or better than the other estimators.

3. Across all three data groups: (1) The three estimators (MLE, LPPR, and PPWM) generally produced comparable L-rmse values when censoring was at or below the 20th percentile when $n = 10$, the 40th percentile when $n = 25$, and the 60th percentile when $n = 50$. The exception was the estimators of X_{10}, whose performance differed at even a lower censoring percentile. (2) At higher censoring, the MLE usually provided the best estimator of quantiles with a nonexceedence probability of 10 and 90 percent, and the interquartile range. The exception is estimators of X_{90} when the shape of the underlying distribution was very different than that of a lognormal distribution, and censoring was at or below the 40th

percentile. (3) "Robust" fill-in methods produced efficient estimators of the mean and standard deviation when used with all three estimators. The MLE generally provided the best estimator of the mean and standard deviation. (4) In general, the LPPR estimators are as good as or better than the PPWM estimators. The LPPR estimators are easier to understand and implement than the PPWM and MLE estimators and thus are recommended for use in practice with medium to large sample sizes and low to moderate censoring.

4. Unlike most other applications of probability-weighted moments (PWMs), the PPWM estimator in this experiment is applied to the logarithms of the data. A log transformation reduces the influence of exceptionally large observations on the estimators of higher moments and quantiles. When the underlying distribution is lognormal, the PPWM, MLE, and LPPR estimators X_{10} and X_{90} are equivalent to real-space estimators for the normal distribution, and the L-rmse is equivalent to a real-space rmse. For normal data the performance of the estimators was very similar for moderate censoring. At higher censoring the MLE performs better than the other estimators.

Acknowledgments. The authors appreciate comments provided by Ed Gilroy, Steve Millard, and Dennis Helsel, who served as reviewers of the manuscript. The authors also express their gratitude toward Richard Vogel, Timothy Cohn, and Jorge Damazio, who also reviewed the manuscript.

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*, Cambridge Univ. Press, New York, 1957.

Cohen, A. C., *Truncated and Censored Samples*, Marcel Dekker, New York, 1991.

Cohen, M. A., and P. B. Ryan, Observations less than the analytical limit of detection: A new approach, *JAPCA, 39*(3), 328–329, 1989.

Cohn, T. A., L. L. DeLong, E. J. Gilroy, R. M. Hirsch, and D. Wells, Estimating constituent loads, *Water Resour. Res., 25*(5), 937–942, 1989.

Condie, R., and G. A. Nix, Modeling of low flow frequency distributions and parameter estimation, paper presented at International Water Resources Symposium: Water For Arid Lands, Teheran, Dec. 8–9, 1975.

David, H. A., *Order Statistics*, John Wiley, New York, 1981.

Fill, H., Improving flood quantile estimates using regional information, Ph.D. thesis, Sch. of Civ. and Environ. Eng., Cornell Univ., Ithaca, N. Y., 1994.

Finney, D. J., On the distribution of a variate whose logarithm is normally distributed, *J. R. Stat. Soc., Ser. B, 7*, 155–161, 1941.

Gilliom, R. J., and D. R. Helsel, Estimation of distributional parameters for censored trace level water quality data, 1, Estimation techniques, *Water Resour. Res., 22*(2), 135–146, 1986.

Gupta, A. K., Estimation of the mean and standard deviation of a normal population from a censored sample, *Biometrika, 39*, 260–273, 1952.

Hammett, K. M., Low-flow frequency analysis for streams in west central Florida, *U.S. Geol. Surv. Water Resour. Invest. Rep., 84-4299*, 1984.

Helsel, D. R., Less than obvious, *Environ. Sci. Technol., 24*(12), 1766–1774, 1990.

Helsel, D. R., and T. A. Cohn, Estimation of descriptive statistics for multiply censored water quality data, *Water Resour. Res., 24*(12), 1997–2004, 1988.

Helsel, D. R., and R. J. Gilliom, Estimation of distributional parameters for censored trace level water quality data, 2, Validation techniques, *Water Resour. Res., 22*(2), 147–155, 1986.

Helsel, D. R., and R. M. Hirsch, *Statistical Methods in Water Resources*, Elsevier Sci., New York, 1992.

Hirsch, R. M., and J. R. Stedinger, Plotting positions for historical floods and their precision, *Water Resour. Res., 23*(4), 715–727, 1987.

Hosking, J. R. M., The theory of probability weighted moments, *Res. Rep. RC12210*, IBM Res. Div., T. J. Watson Res. Cent., Yorktown Heights, N.Y., April 3, 1989.

Hosking, J. R. M., and J. R. Wallis, Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics, 29*(3), 339–349, 1987.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, Estimation of the generalized extreme-value distribution by the method of probability weighted moments, *Technometrics, 27*(3), 251–261, 1985.

Joiner, B. L., and J. R. Rosenblatt, Some properties of the range in samples from Tukey's symmetric lambda distributions, *JASA J. Am. Stat. Assoc., 66*, 394–399, 1971.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee, *The Theory and Practice of Econometrics*, chap. 19, pp. 797–821, John Wiley, New York, 1985.

Kroll, C. N., Censored data analyses in water resources, Ph.D. thesis, Sch. of Civ. and Environ. Eng., Cornell Univ., Ithaca, N. Y., 1996.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis, Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles, *Water Resour. Res., 15*(5), 1055–1064, 1979.

Liu, S., and J. R. Stedinger, Low stream flow frequency analysis with ordinary and Tobit regression, paper presented at 18th Annual Conference and Symposium of the ASCE: Water Resources Planning and Management and Urban Water Resources, Am. Soc. of Civ. Eng., New Orleans, La., May 20–22, 1991.

Newman, M. C., P. M. Dixon, B. B. Looney, and J. E. Pinder, Estimating mean and variance for environmental samples with below detection limit observations, *Water Resour. Bull., 25*(4), 905–916, 1989.

Stedinger, J. R., and T. A. Cohn, Flood frequency analysis with historical and paleoflood information, *Water Resour. Res., 22*(5), 785–793, 1986.

Stedinger, J. R., R. M. Vogel, and E. Georgiou, Frequency analysis of extreme events, in *Handbook of Hydrology: Frequency Analysis of Extreme Events*, chap. 18, pp. 18.1–18.66, McGraw-Hill, New York, 1993.

Tasker, G. D., A comparison of methods for estimating low flow characteristics of streams, *Water Resour. Bull., 23*(6), 1077–1083, 1987.

Tasker, G. D., Regionalization of low flow characteristics using logistic and GLS regression, in *Proceedings of the Baltimore Symposium, New Directions for Surface Water Modeling*, pp. 323–331, Int. Assoc. of Hydrol. Sci., Gentbrugge, Belgium, 1989.

Vogel, R. M., and C. N. Kroll, Low-flow frequency analysis using probability-plot correlation coefficients, *J. Water Resour. Plann. Manage., 115*(3), 338–357, 1989.

Wang, Q. J., Estimation of the GEV distribution from censored samples by method of partial probability weighted moments, *J. Hydrol., 120*, 103–114, 1990.

C. N. Kroll and J. R. Stedinger, School of Civil and Environmental Engineering, Hollister Hall, Cornell University, Ithaca, NY, 14853-3501. (e-mail: ck22@cornell.edu; jrs5@cornell.edu)

(Received January 23, 1995; revised October 19, 1995; accepted October 27, 1995.)