

# Maximum Likelihood Estimators for Quantized Data

S. Rocky Durrans, Associate Professor, Department of Civil and Environmental Engineering  
The University of Alabama, Box 870205, Tuscaloosa, AL 35487-0205

## ***Abstract***

*Measured and archived data for various hydrological and meteorological variables are sometimes truncated or rounded (hereinafter, “quantized”). Quantizing may occur due to the characteristics of a measuring instrument, or from post-processing and quality control of raw data before archival. Ignorance of the nature of a quantizing process can lead to statistical estimation errors, whose magnitudes depend on the type of quantizing (truncation or rounding), on the resolution at which quantized data are reported, on the dispersion (variance or standard deviation) of the data, and on the relative number of zeroes that may be present in a database. This paper presents a maximum likelihood approach to dealing with quantized data, and illustrates the magnitudes of the errors that can arise in parameter and quantile estimates when the effects of data quantizing are ignored. This paper deals specifically with the exponential distribution for demonstration of the errors, but its ideas and results are easily extended to other distributions that may be of interest. Remarks are also made with regard to moment-based estimation approaches.*

## **Introduction**

Measured and archived data for various hydrological and meteorological variables are sometimes truncated or rounded (hereinafter, “quantized”). Quantizing may occur due to the characteristics of a measuring instrument, or from post-processing and quality control of raw data before archival. For instance, many if not most of the precipitation amounts archived by the National Climatic Data Center (NCDC) since about 1972 are reported to only the nearest 2.5 mm (0.1 in.). While that resolution is more than adequate for many data applications, it must be considered to be quite coarse for frequency analyses of small precipitation depths. Small depths arise in frequency analyses for short storm durations, and also arise in connection with design and implementation of best management practices (BMPs) for control of pollutants in urban storm water runoff.

Ignorance of the nature of a quantizing process can lead to statistical estimation errors. The magnitudes of the errors depend on the type of quantizing (truncation or rounding), on the resolution at which quantized data are reported, on the dispersion (variance or standard deviation) of the data, and on the relative numbers of zeroes that may be present in a database. For a fixed data resolution, errors can be expected to be more significant for random processes exhibiting a small amount of dispersion than a large amount. In the context of rainfall analysis, for example, errors can be expected to be greater in arid or semiarid regions (where recorded rainfall depths are typically small) than in wet regions, and can be expected to be more significant for short storm durations than for long ones (again, because of the relative magnitudes of the depths). Zeroes can present particularly difficult problems, especially for intermittent processes such as precipitation. In particular, given a time series of recorded precipitation

amounts, it is difficult to tell whether one or more zeroes preceding or trailing a wet episode are true zeroes, or whether they are zeroes only because of the quantizing process. Problems related to this have been addressed by Kroll and Stedinger (1996) and Durrans et al. (1999).

Given the preponderance of recent and ongoing precipitation frequency studies, both within and external to the National Weather Service (see Durrans and Brown, 2001, for a summary of them), and given the ubiquity of quantized rainfall records utilized in those studies, this paper presents a statistical estimation approach that can be applied to deal with those data. It should be noted, however, that the problems posed by quantizing are not limited to applications involving frequency analyses of annual or partial duration series of maximums. Frequency analyses of any variable, such as rainfall totals on wet days, are instances in which the ideas and methods presented herein may be relevant.

The estimation approach presented here is a maximum likelihood (MLE) one, though remarks are also forwarded with respect to moment-based estimation methods. The estimation approach presented here is related to Sheppard's corrections for grouping (see Kendall and Stuart, 1963), and to similar work by Lindley (1950). It is also related to MLE methods for treatment of censored data, such as those described by Leese (1973) and Stedinger and Cohn (1986). This paper deals primarily with the exponential distribution for demonstration of the errors that may arise in parameter and quantile estimates, but the underlying ideas are easily extended to other distributions. This paper does not address the problem of discrimination of true zeroes from quantized ones, as that is a more difficult problem deserving of additional research. Where zeroes must be dealt with in this paper, it is assumed that their presence is due to quantizing, and that an adequate means has been applied to distinguish them from true zeroes.

## MLEs for Quantized Data

**General Remarks Regarding MLEs.** Given a vector of data  $\mathbf{X} = \{X_1, X_2 \dots X_n\}'$ , and a distributional model  $F$  from which the data are hypothesized to have been drawn, the joint probability of having observed the vector of data, given the specified form of the model, is  $\mathbf{x} = P(\mathbf{X}|F)$ . If the data are independent then

$$\mathbf{x} = \prod_{i=1}^n dF_X(x_i) = \prod_{i=1}^n f_X(x_i) dx \tag{1}$$

Here  $F_X(x)$  and  $f_X(x)$ , respectively, represent the cumulative distribution function (cdf) and probability density function (pdf) of the model, and  $x_i$  represents the  $i$ -th observed sample value.

Estimates of the parameters of the hypothesized model may be obtained by maximizing  $\mathbf{x}$  with respect to each unknown parameter.

An alternative form of (1) may be written as

$$L = \frac{\mathbf{x}}{(dx)^n} = \prod_{i=1}^n f_X(x_i) \tag{2}$$

This is known as the likelihood function. Note that it differs from (1) by the factor  $1/(dx)^n$ . Note also that it does not have the non-dimensional units of probability; it has instead the units of  $1/w$ , where  $w$  denotes the units associated with the random variable  $X$ . Either (1) or (2) may be used to estimate the parameters of a chosen model, as the location of a stationary point (i.e. a maximum) on the likelihood or joint probability functions is unaffected by the multiplier  $(dx)^n$ .

In practice, it is often the log-likelihood function that is maximized rather than the likelihood function. That is, one maximizes the quantity

$$\ln L = \ln \prod_{i=1}^n f_X(x_i) = \sum_{i=1}^n \ln f_X(x_i) \quad (3)$$

This is done primarily for mathematical convenience. Because of the monotonicity of the logarithmic transform, this also does not affect estimated parameter values.

The expressions (1) through (3), and especially (2) and (3), are well known and normally would not need to be expressed in a paper dealing with statistical parameter estimation. They are presented here for the purpose of distinguishing between the joint probability function defined by (1), and the likelihood function forms of it that are usually applied in practice [Eqns. (2) and (3)]. When data are not quantized, or when the resolution  $\Delta x$  at which data are quantized is small, the forms (2) and (3) are adequate. However, when the effects of quantizing are pronounced, the joint probability function (1), or its logarithmic transform, should be used.

**MLEs for Truncated Data.** When data are truncated, the value of a random variable falling within an interval  $[x, x + \Delta x)$  is reported as the lower limit of that interval, or as  $x$ . For example, if  $\Delta x = 2.5$  mm (0.1 in.), an actual rainfall depth of 4.3 mm (0.17 in.) would be reported as 2.5 mm (0.1 in.). The probability of the true value falling in the interval is  $F_X(x + \Delta x) - F_X(x)$ . By definition, the joint probability function is the product of these probabilities when data are independent and identically distributed. Using (1), it is expressed as

$$\mathbf{x}_T = \prod_{i=1}^n dF(x_i) = \prod_{i=1}^n \int_{x_i}^{x_i + \Delta x} f_X(x) dx = \prod_{i=1}^n \{F_X(x_i + \Delta x) - F_X(x_i)\} \quad (4)$$

Note that the subscript on  $\mathbf{x}_T$  is used to emphasize that this form of the joint probability function applies to truncated data.

A logarithmically-transformed version of (4) is often more convenient than is (4) itself. It is expressed as

$$\ln \mathbf{x}_T = \sum_{i=1}^n \ln \int_{x_i}^{x_i + \Delta x} f_X(x) dx = \sum_{i=1}^n \ln \{F_X(x_i + \Delta x) - F_X(x_i)\} \quad (5)$$

Again, because of the monotonicity of the logarithmic transform, either (4) or (5) may be maximized to estimate the parameters of a model on the basis of an observed set of data.

**MLEs for Rounded Data.** When data are rounded, the value of a random variable falling within an interval  $[x - \frac{1}{2}\Delta x, x + \frac{1}{2}\Delta x)$  is reported as the central value of that interval, or as  $x$ . For example, if  $\Delta x = 2.5$  mm (0.1 in.), an actual rainfall depth of 4.3 mm (0.17 in.) would be reported as 5.0 mm (0.2 in.). The probability of the true value falling in the interval is  $F_X(x + \frac{1}{2}\Delta x) - F_X(x - \frac{1}{2}\Delta x)$ . Thus, the joint probability function for this case is

$$\mathbf{x}_R = \prod_{i=1}^n dF(x_i) = \prod_{i=1}^n \int_{x_i - \Delta x/2}^{x_i + \Delta x/2} f_X(x) dx = \prod_{i=1}^n \left\{ F_X \left( x_i + \frac{\Delta x}{2} \right) - F_X \left( x_i - \frac{\Delta x}{2} \right) \right\} \quad (6)$$

The subscript  $R$  is used to emphasize that this form applies to rounded data. The log-joint probability function for rounded data is expressed as

$$\ln \mathbf{x}_R = \sum_{i=1}^n \ln \int_{x_i - \Delta x/2}^{x_i + \Delta x/2} f_X(x) dx = \sum_{i=1}^n \ln \left\{ F_X \left( x_i + \frac{\Delta x}{2} \right) - F_X \left( x_i - \frac{\Delta x}{2} \right) \right\} \quad (7)$$

Care needs to be exercised in an application of either (6) or (7) when rounding causes one or more data observations  $x_i$  to take on the value  $x_0$ , where  $x_0$  is the lower bound of a distributional model for the data. In such cases, alternative (but formally equivalent) forms of (6) and (7) are

$$\mathbf{x}_R = \prod_{i=1}^{n_0} F_X \left( x_0 + \frac{\Delta x}{2} \right) \prod_{j=1}^{n_1} \left\{ F_X \left( x_j + \frac{\Delta x}{2} \right) - F_X \left( x_j - \frac{\Delta x}{2} \right) \right\} \quad (8)$$

$$\ln \mathbf{x}_R = n_0 \ln \left\{ F_X \left( x_0 + \frac{\Delta x}{2} \right) \right\} + \sum_{j=1}^{n_1} \ln \left\{ F_X \left( x_j + \frac{\Delta x}{2} \right) - F_X \left( x_j - \frac{\Delta x}{2} \right) \right\} \quad (9)$$

In these expressions,  $n_0$  is the number of observations equal to  $x_0$ , and  $n_1 = n - n_0$  is the remaining number of larger observations. These alternative expressions take note of the fact that rounding may take place in the half-interval  $[x_0, x_0 + \frac{1}{2}\Delta x)$ , and that in such cases  $F_X(x_0 - \frac{1}{2}\Delta x) = 0$ . Note that alternative expressions analogous to these are not necessary for truncated data.

### Application to the Exponential Distribution

Applications of the joint probability functions for quantized data to specific distributional models are straightforward. This paper deals with the exponential distribution for illustrative purposes. The exponential distribution has been chosen for application herein as it is relatively tractable from a mathematical point of view. That distribution is also applied in many practical applications, and has been used extensively for modeling depths of precipitation on wet days. The cdf of the exponential distribution is of the form

$$F_X(x) = 1 - e^{-Ix} \quad (10)$$

where  $I > 0$  is a parameter and  $X \geq 0$ . The mean and standard deviation of this distribution are equal, and are  $\mathbf{m} = \mathbf{s} = 1/I$ . The coefficient of skewness is independent of  $I$ , and is equal to 2.

When data are not quantized by truncation or rounding, application of the log-likelihood function (3) yields an estimator for the parameter  $I$  of the exponential model as

$$I = \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^{-1} = \frac{1}{\bar{x}} \quad (11)$$

where  $\bar{x}$  is the arithmetic mean of the sample data.

**Estimators for Quantized Data.** When data are truncated with a resolution of  $\Delta x$ , an application of (5) to the exponential distribution yields

$$\ln \mathbf{x}_T = -I \sum_{i=1}^n x_i + n \ln(1 - e^{-I\Delta x}) \quad (12)$$

The derivative of this expression with respect to  $I$ , when equated to zero, yields an estimator for the parameter as

$$I_T = \frac{1}{\Delta x} \ln\left(1 + \frac{\Delta x}{\bar{x}}\right) \quad (13)$$

where  $\bar{x}$  is the arithmetic mean of the sample of truncated data. The subscript  $T$  is used for this estimator to distinguish it from the estimator expressed by (9). Note that the limit of  $I_T$ , as  $\Delta x$  approaches zero, is equal to  $I$  as given by (9).

The exponential distribution, because of its lower bound of zero, represents a case where rounding may cause one or more data values to assume the value of zero. Thus, for rounded data with a resolution of  $\Delta x$ , an application of (9) yields

$$\ln \mathbf{x}_R = n_0 \ln[1 - e^{-I\Delta x/2}] + n_1 \ln[e^{I\Delta x/2} - e^{-I\Delta x/2}] - I \sum_{j=1}^{n_1} x_j \quad (14)$$

The parameter estimator in this case is implicit and requires solution of the expression

$$\frac{(n_0 - n_1) - (n_0 + n_1)e^{-I\Delta x} + n_1(e^{I\Delta x/2} + e^{-I\Delta x/2})}{e^{I\Delta x/2} - e^{-I\Delta x/2} + e^{-I\Delta x} - 1} - \frac{2}{\Delta x} \sum_{j=1}^{n_1} x_j = 0 \quad (15)$$

Hereinafter, the estimate of  $I$  satisfying this expression will be denoted by  $I_R$ .

**Qualitative Properties of Parameter and Quantile Estimators.** Because the estimators  $I_T$  and  $I$  are expressible in explicit forms, their ratio may be used to illustrate the effects of truncation on parameter estimation for the exponential model. That ratio is

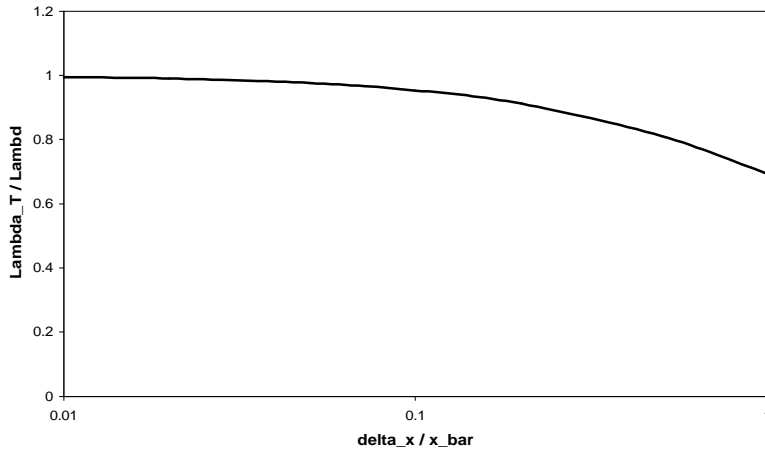
$$\frac{I_T}{I} = \frac{\bar{x}}{\Delta x} \ln\left(1 + \frac{\Delta x}{\bar{x}}\right) \quad (16)$$

A graph of this relationship is illustrated in Figure 1. It may be seen that the ratio is always positive but smaller than unity, and that it approaches unity as  $\Delta x$  approaches zero. In other words,  $I_T$  is a fraction of  $I$ , as expected. Figure 1, or (16), may be used to “correct” an estimate of  $I$  incorrectly computed using (11) for a sample of truncated data. Alternatively, one could apply (13) to compute the correct estimate directly.

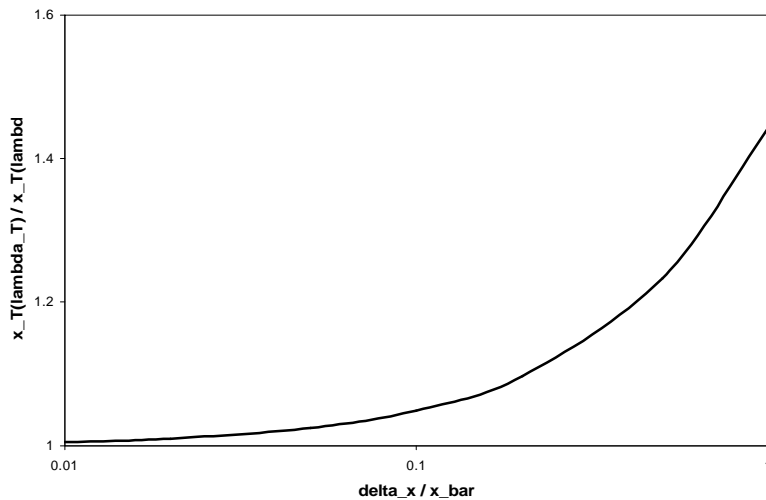
That  $I_T$  is a fraction of  $I$  could have been deduced qualitatively on the basis of statistical reasoning, for it implies that an inferred population mean  $m_T = 1/I_T$  is larger than an inferred population mean  $m = 1/I$ . This makes sense because a sample mean computed on the basis of truncated data would be lower than the sample mean that would have been computed had the data not been truncated.

Similar reasoning can be applied to the case of rounding, where the ratio  $I_R / I$  cannot be expressed explicitly for the exponential model. In this case, a sample mean computed on the basis of rounded data would be larger than the sample mean that would have been computed had the data not been rounded. Thus,  $I_R$  would be larger than  $I$ , and the inferred population mean would be smaller. This again makes sense because of the reverse- $J$  shape of the exponential density function, which places more probability in an interval  $(x - 1/2\Delta x, x)$  than in an interval  $(x, x + 1/2\Delta x)$ .

The effects of data truncation and rounding on quantile estimates may be reasoned as follows. Figure 2 shows the ratio  $X_T(I_T)/X_T(I)$ , where  $X_T(I_T) = (\ln T)/I_T$  is the estimator of the  $T$ -year quantile  $X_T$  based on  $I_T$ , and  $X_T(I) = (\ln T)/I$  is the corresponding estimator based on  $I$ . The ratio is  $X_T(I_T)/X_T(I) = I/I_T$ , and thus the relationship graphed in Figure 2 is just the reciprocal of that shown in Figure 1. Note that the relationship shown in Figure 2 is independent of the return period  $T$  associated with a quantile estimate.



**Figure 1.** The ratio  $I_T/I$  as a function of  $\Delta x/m = \Delta x/s$ . Note that quantizing errors may be large if  $\Delta x > 0.1s$ .



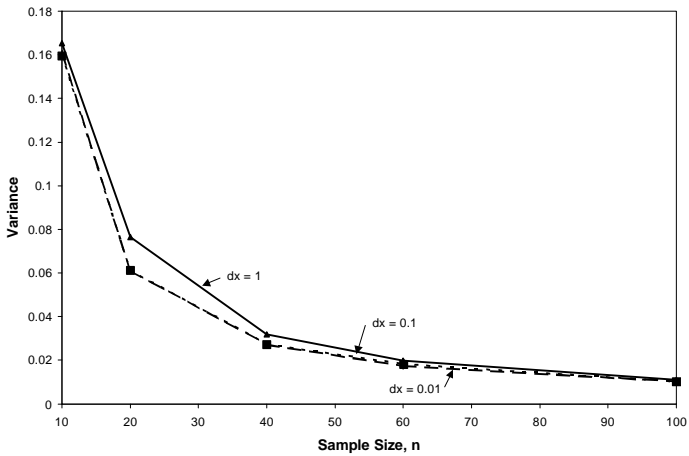
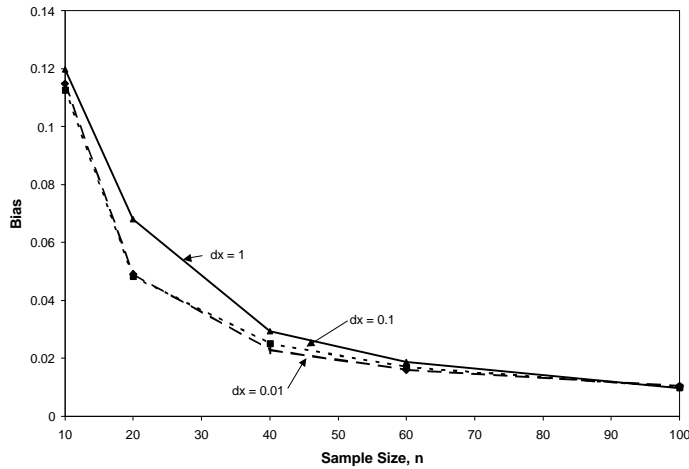
**Figure 2.** The ratio  $X_T(I_T)/X_T(I)$  as a function of  $\Delta x/m = \Delta x/s$ . Note that quantizing errors may be large if  $\Delta x > 0.1s$ .

**Quantitative Properties of Parameter and Quantile Estimators.** Equation (16), and the statements following it based on statistical reasoning, say little about the biases and variances of alternative parameter and quantile estimators. It is these quantitative properties, rather than the more qualitative ones just described, that provide support for the use of an estimator.

Herein, a demonstration of the biases and variances associated with parameter estimators is limited to the estimators (11) and (13), as they represent alternative approaches to parameter estimation when data are truncated. It is intended to show that (13) is superior to (11), as (13) acknowledges the truncation process. The statistical qualities of parameter estimates obtained using (15) versus (11) are not demonstrated here, as statistical reasoning clearly indicates that the effects of data quantizing are more significant for truncation than for rounding. In a similar vein, biases and variances associated with quantile estimators also are not demonstrated herein. By showing that (13) yields smaller biases and variances in parameter estimates than does (11), it follows that quantile estimates derived from use of (13) would also have smaller biases and variances than those derived from use of (11).

Figure 3 illustrates the bias and variance of the estimator  $I_T$  expressed by (13). Those results have been obtained by Monte Carlo generation of samples from an exponential population with  $I = 1$ . For each sample size  $n$ , randomly generated sample values were truncated at a resolution of  $\Delta x$ . For each combination of  $n$  and  $\Delta x$  shown by data points in the figures, 10,000 replicate samples were generated.

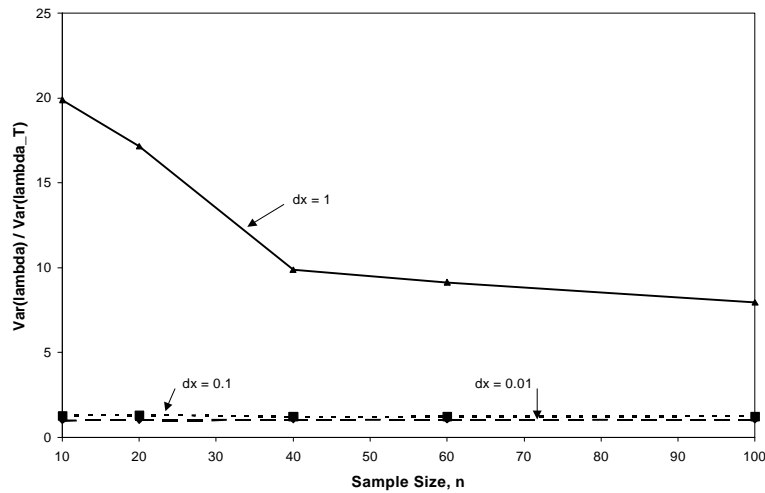
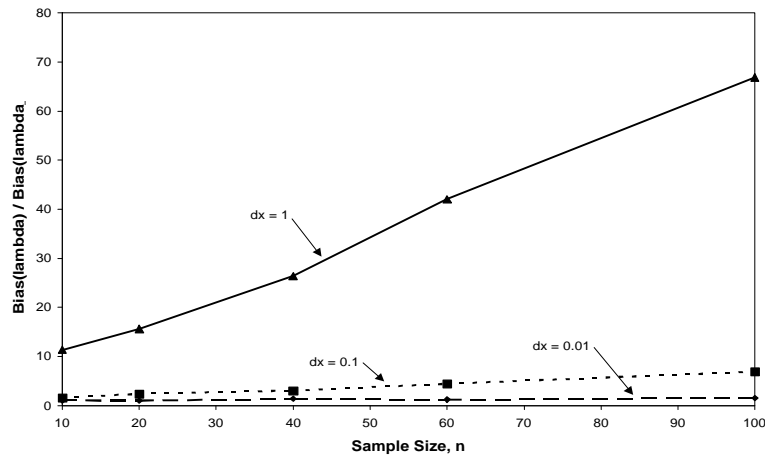
Figure 3 clearly shows that the (always positive) bias of the estimator decreases with increasing sample size, and that it is only slightly dependent on the data resolution  $\Delta x$ . The bias for an extremely large  $\Delta x = 1$  (when  $\Delta x = 1$ , it is equal to the standard deviation of the population, as the population has  $I = 1$  and  $s = 1/I$ ) is only marginally larger than that when  $\Delta x = 0.01$ . The variance of the estimator decreases with increasing sample size, and it too is only marginally dependent on the data resolution. These insensitivities of the bias and variance to  $\Delta x$  are very nice properties of the estimator, as they demonstrate that it can do nearly as well with



**Figure 3.** Bias (top) and variance (bottom) of the estimator  $I_T$  as a function of sample size and data resolution. The population is exponential with  $I = 1$ .

very coarsely resolved data as it can with finely resolved data.

Results shown in Figure 4 also have been obtained via simulation using an exponential population with  $I = 1$ . The top panel shows the ratio of the bias of  $I$  to the bias of  $I_T$  when the estimators are applied to samples of size  $n$  with truncated sample values. Note that the bias ratio is essentially constant and equal to unity when  $\Delta x$  is small. This makes sense as (13) converges to (11) as  $\Delta x$  decreases. For a data resolution equal to one-tenth of the population standard deviation (i.e. when  $\Delta x = 0.1$ ), the bias ratio varies from about 1.5 to about 6 as the sample size increases from 10 to 100. This illustrates that the bias of the estimator (13) can be significantly smaller than that of (11) in real applications involving truncated data, especially when sample sizes are moderate to large. When  $\Delta x$  increases to the very large value of unity, the bias of (13) varies from about 1/10 to about 1/65 of that of (11), depending on the sample size.



**Figure 4.** Ratios of biases (top) and variances (bottom) of estimators for  $I$  and  $I_T$ . The estimator  $I_T$  is always less biased and less variable than is the estimator  $I$ .

The lower panel of Figure 4 shows that the ratio of the variances of the estimators (11) and (13) does not depend as strongly on  $\Delta x$  as does the ratio of the biases. Nevertheless, it continues to illustrate the superiority of the truncated data estimator (13). The variance of estimates obtained via (13) is always smaller than that of estimates obtained via (11).

### Moment-Based Methods

There are a number of moment-based methods for parameter estimation, including classical product-moments, probability-weighted moments (PWMs) (Greenwood et al., 1979), and  $L$ -moments (Hosking, 1990). These methods, like MLEs, can be adapted to deal with quantized data, but the mathematics of the adaptations tend to be more cumbersome as they usually require evaluations of the sums of infinite series to express the population moments. The series must certainly converge if the ordinary population

moments exist, but evaluations of their sums are difficult and are not attempted here. Evaluations of the sums are left for future work.

## Conclusions

Consideration of the effects of data quantizing in the MLE parameter estimation process can lead to considerable gains. In the case of data truncation, bias is reduced considerably, and variance is also reduced. The truncated MLE estimator (13) is no more difficult to apply than is the commonly used expression (11), and is consistent with (11). Dealing with rounded data is more problematic from the parameter estimation point of view, but is not insurmountable. It simply involves numerical iteration to obtain the root of (15).

The MLE approach to estimation has been chosen for presentation in this paper for mathematical reasons. Estimators based on moment-based approaches, for most probability distributions of practical interest, would involve evaluations of the sums of infinite series. Those sums should certainly exist if the ordinary raw moments of the distribution exist, but are difficult to evaluate. Future work should address those types of estimators in more detail.

## References

Durrans, S.R., T.B.M.J. Ouarda, P.F. Rasmussen, and B. Bobee, Treatment of zeroes in tail modeling of low-flows, *Journal of Hydrologic Engineering*, ASCE, 4(1):19-27, 1999.

Durrans, S.R., and P.A. Brown, Estimation and web-based dissemination of extreme rainfall information, *Transportation Research Record*, in press, 2001.

Greenwood, J.A., J.M. Landwehr, N.C. Matalas, and J.R. Wallis, Probability-weighted moments: Definition and relation to parameters of distributions expressible in inverse form, *Water Resources Research*, 15(5):1049-1054, 1979.

Hosking, J.R.M., *L*-moments: Analysis and estimation of distributions using linear combinations of order statistics, *Journal of the Royal Statistical Society, Ser. B*, 52(2):105-124, 1990.

Kendall, M.G., and A. Stuart, *The advanced theory of statistics*, vol. 1, Charles Griffin, London, 1963.

Kroll, C.N., and J.R. Stedinger, Estimation of moments and quantiles using censored data, *Water Resources Research*, 32(4):1005-1012, 1996.

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

Lindley, D.V., Grouping corrections and maximum likelihood equations, *Proceedings of the Cambridge Philosophical Society*, 46:106, 1950.

Stedinger, J.R., and T.A. Cohn, Flood frequency analysis with historical and paleoflood information, *Water Resources Research*, 22(5):785-793, 1986.