International Seminar on Hydrology of Extremes 1-3 December, 1988
Roorkee (India)

APPLICATION OF LP3 DISTRIBUTION TO CENSORED SAMPLES

Huynh Ngoc Phien
Associate Professor

Yang Jung-Hsiung
Graduate Student

Asian Institute of Technology
P.O. Box 2754, Bangkok 10501, Thailand.

SYNOPSIS

The effects of the type-1 censoring on the maximum likelihood estimators based on samples drawn from the Log Pearson Type-3 (LP3) distribution were analyzed in this study. For this purpose, all the equations needed in obtaining the maximum likelihood estimators and their asymptotic variances-covariances derived. From Monte Carlo experiments, it was found that censoring may reduce the bias in estimating the T-year event but increases its variance. It was also found that the type-1 censoring can be effectively used in dealing with samples containing outliers.

1.0 INTRODUCTION

Flood frequency analysis deals with maximum values of river flows. Because of the difficulty and danger in the measurement of big flood discharges, their values are obtained mostly by extrapolating the rating curve established at the station concerned. However, such a rating curve is normally established from a small range of measured values of the discharge and water level (stage), consequently, it may no longer valid for very high values. Thus, very high discharges (beyond the range of the established rating curve) are known to occur, but their individual values are not known with the same reliable degree as the other normal floods. As such, they should be considered to be right censored.

In arid areas, flood data may consist of zero or extremely low values. These should be well treated as being left censored.

This study applies the type-1 censoring to the case where annual flood data are assumed to follow the log Pearson type-3 (LP3) distribution. The effects of censoring on the maximum likelihood estimators are investigated by means of Monte Carlo experiments. It is also proposed that outliers be treated as censored values.

2.0 METHODOLOGY

2.1 The LP3 Density Function

The LP3 distribution has the following density function [5]:

$$f(x) = f(x;a,b,c) = \frac{1}{|a|x\Gamma(b)} \left(\frac{\ln x - c}{a}\right)^{-1} \exp\left[-\frac{\ln x - c}{a}\right] \qquad (1)$$

where a,b,c are the scale, shape and location parameters, respectively, and Γ is the gamma function. The random variable Y defined by $y = \ln X$ has a three-parameter gamma (or Pearson type-3) distribution.

2.2 Censoring

Consider a random sample of size N, where m values on the left of the lower point X1, k values on the right of the upper point Xu are only known to exist (but not their individual values), and n = N-m-k values in the middle are observed. If X1 and Xu are fixed, then the sample is censored according to type-1. In this case, m, n and k are random variables. If m and k are fixed, then the sample is censored according to type-2. In such a case, X1 and Xu are random variables. As mentioned earlier, the type-1 censoring is considered because X1 and Xu are often known (fixed) in flood frequency analysis.

2.3 Maximum Likelihood Equations

The method of maximum likelihood (ML) is readily applicable to estimate the parameters of the LP3 distribution from censored samples. The likelihood function of such a sample is given by

$$[N!/(m!k!)]\left[\int_{-\infty}^{X1} f(x)dx\right]\left[\prod_{i=1}^{n} f(x_i)\right]\left[\int_{Xu}^{\infty} f(x)dx\right]$$

where n = N-m-k is the number of uncensored points.
The log-likelihood function is then:

$$L = \ln(N!) - \ln(m!) - \ln(k!) + m\ln p + \sum_{1}^{n} \ln f(x) + k\ln q$$

where

$$\begin{aligned} p &= \text{Prob}(X \le X1) = I(W1,b)\ ,\ q = \text{Prob}(X > Xu) = 1 - I(Wu,b) \quad \text{for } a>0 \\ p &= 1 - I(W1,b)\ ,\ q = I(Wu,b) \quad \text{for } a<0 \end{aligned} \qquad (2)$$

In these equations,

$$w = (\ln x - c)/a$$

$$I(w,b) = \frac{1}{\Gamma(b)} \int_{u}^{w} t^{b-1} e^{-t}\, dt \qquad (3)$$

By setting the partial derivatives of L with respect to a,b and c equal to zero,

one obtains the ML equations:

$$\begin{aligned} H1 &= (1/a)(P+P1+Pu) = 0 \\ H2 &= Q+Q1+Qu = 0 \\ H3 &= (1/a)(R+R1+Ru) = 0 \end{aligned} \qquad (4)$$

where

$$\begin{aligned} P &= (n/a)(\bar{y}-c-ab) \\ P1 &= -sign(a)*mW1g(W1)/p \\ Pu &= sign(a)*k*Wu*g(Wu)/q \end{aligned} \qquad (5)$$

$$\begin{aligned} Q &= \sum_{1}^{n} lnw-n\Psi(b) \\ Q1 &= sign(a)*mI'(W1,b)/p \\ Qu &= -sign(a)*k+I'(Wu,b)/q \end{aligned} \qquad (6)$$

$$\begin{aligned} R &= n-(b-1)\sum_{1}^{n}(1/W) \\ R1 &= -sign(a)*m*g(W1)/p \\ Ru &= -sign(a)*k*g(Wu)/q \end{aligned} \qquad (7)$$

In these above equations,

$$g(w) = \frac{1}{\Gamma(b)} w^{b-1} e^{-w}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\Psi(b) = \frac{d}{db} ln\Gamma(b) = \text{digamma function of b [1]}$$

and I'(w,b) denotes the derivative of the imcomplete gamma function I(w,b) with respect to b (Moore [3]).

The ML estimates of a,b,c are obtained by maximizing L. In practice, they are obtained by solving the system expressed by eq. 4 and by imposing the conditions for a local maximum. In this connection, Newton's method (see Rice, [7]) can be used.

2.4 Asymptotic Variances-Covariances of ML Estimators

The asymptotic variances-covariances of the estimators of a,b and c can be obtained by inverting the Fisher information matrix:

$$F = -\begin{bmatrix} E(\partial^2 L/\partial a^2) & E(\partial^2 L/\partial a \partial b) & E(\partial^2 L/\partial a \partial c) \\ & E(\partial^2 L/\partial b^2) & E(\partial^2 L/\partial b \partial c) \\ & & E(\partial^2 L/\partial c^2) \end{bmatrix}$$

where E stands for the mathematical expectation operator. It is quite tedious to evaluate the expections involved, hence only their expressions are given here.

From Leese [2], one can write:

$E(n) = N(1-p-q)$
$E(m) = Np$
$E(k) = Nq$

where p and q are defined by

$p = \text{Prob}(X \leq Xl)$
$q = \text{Prob}(X \geq Xu)$

(a) Censored from below:

Let $E_{11}(W) = [1/(1-p)] \int_{Wl}^{\infty} wg(W)\,dW$, $E_{12}(W) = (1/p) \int_{0}^{Wl} wg(W)\,dW$

then $pE_{11}(W) + (1-p)E_{12}(W) = E(w) = b$

$$pE_{11}(W) = \frac{b}{b\Gamma(b)} \int_0^{Wl} W^b e^{-w} dw = bI(Wl,b+1)$$

$$E_{12}(W) = b[1-I(Wl,b+1)]/(1-p)$$

$$pE_{11}(W^{-1}) + (1-p)E_{12}(W^{-1}) = E(w^{-1})$$

$$E(W^{-1}) = \frac{1}{(b-1)\Gamma(b-1)} \int_0^{\infty} W^{(b-1)-1} e^{-w} dw = \frac{1}{b-1}$$

$$pE_{11}(W^{-1}) = \frac{1}{b\Gamma(b)} I(Wl,b-1)$$

$$E_{12}(W^{-1}) = \frac{1-I(Wl,b-1)}{(b-1)(1-p)}$$

$$E_{12}(W^{-2}) = \frac{1-I(W1,b-2)}{(b-2)(b-1)(1-p)}$$

$$E_{12}(\ln W) = \frac{1-I'(W1,b)+I(W1,b)\Psi(b)}{1-p}$$

(b) Censored from above:

Define $(1-q)E_{21}(W) = \frac{1}{\Gamma(b)}\int_0^{Wu} W^b e^{-w}dw = bI(Wu,b+1)$

then $E_{21}(W) = bI(Wu,b+1)/(1-q)$

In the same way:

$$(1-q)E_{21}(W^{-1}) = \frac{1}{\Gamma(b)}\int_0^{Wu} W^{(b-1)-1}e^{-w}dw$$

or $E_{21}(W^{-1}) = I(Wu,b-1)/[(b-1)(1-q)]$

Similarly $E_{21}(W^{-2}) = I(Wu,b-1)/[(b-2)(b-1)(1-q)]$

$$E_{21}(\ln W) = \frac{I'(Wu,b)-I(Wu,b)\Psi(b)}{1-q}$$

(c) Double censoring:

$$E_3(W) = \frac{b[I(Wu,b+1)-I(W1,b+1)]}{1-p-q}$$

$$E_3(W^{-1}) = \frac{I(Wu,b-1)-I(W1,b-1)}{(b-1)(1-p-q)}$$

$$E_3(W^{-2}) = \frac{I(Wu,b-2)-I(W1,b-2)}{(b-2)(b-1)(1-p-q)}$$

$$E_3(\ln W) = \frac{I'(Wu,b)-I'(W1,b)=(I(Wu,b)-I(W1,b))\Psi(b)}{1-p-q}$$

where the $E_{12}(.)$, $E_{21}(.)$ and $E_3(.)$ denote the expected value of the left, right and double censoring, respectively.

3.0 MONTE CARLO EXPERIMENTS

To investigate the effects of censoring on the maximum likelihood (ML) estimators, Monte Carlo experiments were conducted. The combined algorithm by Phien and Ruksasilp [6] was used to produce the one-parameter gamma variable w, then the LP3 variable is obtained by taking

$$X = \exp(c+aW) \tag{8}$$

Instead of investigating the effects of cnesoring on the ML estimators of a,b and c, it is more convenient to concentrate on the T-year event:

$$\text{Prob}(X \leq X_T) = 1-1/T \tag{9}$$

where T was taken equal to 100, 500 and 1000 years in this work.

3.1 Censoring Points

The censoring points Xl and Xu are determined by fixing the values of the following probabilities:

$$p = \text{Prob}(X \leq Xl) \ , \ 1-q = \text{Prob}(X \leq Xu) \tag{10}$$

With given values of a,b and c, and given values of p and q, Xl and Xu can be computed using the algorithm developed by Phien [4], and are then fixed.

3.2 Outliers

The study also tries to explore the type-1 censoring in dealing with extremely large values (upper outliers). Obviously, their given values are questionable, and these should be censored. In Monte Carlo experiments carried out in this work, the outlier was generated as follows. Let Wu denote the one-parameter gamma quantile correspoinding to a high probability of non-exceedance p. For a chosen sample size n, one generates n one-parameter gamma variables, all forced to be less than Wu. Let

$$Wo = \alpha \max\{w_1, w_2, \ldots w_n\}$$

where α is selected to make WO $\geq$ 1.2 Wu.
An outlier is then obtained as

$$Xo = \exp(aWo+c)$$

Three situation were considered here:

(i) The outlier was included in the sample, i.e. N = n+1.
(ii) The outlier was excluded, i.e. N = n, and
(iii) The outlier was censored with right censoring at Xu = exp(aWu+c).

In this case N = n+1 , m = 0 and k = 1.

3.3 Performance Indices

As mentioned earlier, the ML estimator of the T-year event X is considered. Let $\hat{X}_T$ denote the ML estimate of X_T, then the relative error is

$$e = 100(X_T - \hat{X}_T)/X_T$$

Let $\hat{a}$,$\hat{b}$ and $\hat{c}$ denote the ML estimators of a,b and c, then

$$\hat{X}_T = \exp(\hat{a}W_T + \hat{c})$$

where W_T is the one-parameter gamma quantile (depending on $\hat{b}$) corresponding to $1-1/T$ for $\hat{a}>0$, and to $1/T$ for $\hat{a}<0$.

When a very large number of replications is used, the relative bias can be obtained as:

$$\text{Bias}(\%) = (1/M) \sum_{i=1}^{M} e_i$$

and the Root Mean Square Error is

$$\text{RMSE}(\%) = [(1/M) \sum_{i=1}^{M} e_i^2]^{\frac{1}{2}}$$

In these equations, M is the number of sequences among 10000 replications where Newton's method used solving eq. 4 is convergent.

4.0 RESULTS AND DISCUSSIONS

4.1 Effects of Censoring

Several sents of values of a,b and c were used in the simulation experiments. Typical results are collected for the case a = 0.05, b = 10.2 and c = 6.0, in Tables 1, 2 and 3, respectively forleft, right and double censoring.

From these tables, it is seen that left censoring may reduce the relative bias and root mean square error (RMSE) of the T-year event estimator (compare the case p = 0.05 in Table 1 with p = q = 0 in Table 3). In terms of the RMSE, the results for left censoring appear to be quite expectable.

(a) When the censoring level is large, the RMSE gets larger. This may be explained by the fact that when p increases, the "actual sample size" n decreases, giving rise to large variance and hence large RMSE.
(b) When the sample size increases, the RMSE decreases.
(c) When T increases, i.e. the event becomes more critical, the RMSE becomes larger.

For right censoring (Table 2) at low level (q = 0.05), the same situation can be observed. Then more irregularities appear such as the fact that when N increases, the RMSE also increases (for $q \geq 0.10$). These irrigularities are believed to be due to a relatively small number of sequences for which Newton's method was convergent.

Table 1 Effects of left censoring on the estimator of the T-year event

Sample size N	Relative bias (%) T = 100	T = 500	T = 1000	Root mean square error (%) T = 100	T = 500	T = 1000
			p = 0.05			
30	0.4	1.2	2.1	17.8	23.4	26.0
50	-1.3	-2.8	-3.6	11.8	16.4	18.4
70	-1.6	-2.9	-3.6	8.9	12.6	14.1
			p = 0.10			
30	-1.0	-3.6	-5.1	18.4	24.8	27.9
50	-2.2	-4.4	-5.5	13.0	17.8	20.0
70	-1.6	-3.4	-4.3	12.2	16.1	17.8
			p = 0.15			
30	-0.5	-3.5	-5.1	20.4	26.8	29.8
50	-1.8	-4.5	-5.8	15.4	20.4	22.8
70	-3.2	-5.9	-7.3	12.2	17.0	19.2
			p = 0.20			
30	2.5	-0.1	-1.5	24.0	31.6	34.7
50	-1.4	-4.3	-5.7	17.0	22.1	24.6
70	-1.3	-3.7	-4.9	15.8	20.4	22.8

$p = Prob(X \leq X1)$

Table 2 Effects of right censoring on the estimator of the T-year event

Sample size N	Relative bias (%) T = 100	T = 500	T = 1000	Root mean square error (%) T = 100	T = 500	T = 1000
			q = 0.05			
30	13.1	16.6	17.9	20.6	24.7	26.3
50	13.4	17.3	18.8	19.1	23.3	24.8
70	13.5	17.5	19.1	18.6	22.7	24.3
			q = 0.10			
30	23.7	29.0	30.9	30.9	36.7	38.8
50	28.1	34.1	35.4	34.7	41.1	43.3
70	28.3	34.5	36.8	34.7	41.1	43.3
			q = 0.15			
30	28.1	34.0	36.2	34.7	41.1	43.4
50	35.6	42.6	45.1	40.8	48.0	50.5
70	40.4	47.9	50.7	44.6	52.4	55.1
			q = 0.20			
30	31.4	37.8	40.1	37.1	44.0	46.4
50	40.8	48.4	51.1	44.7	52.5	55.2
70	45.7	53.9	56.8	48.3	56.7	59.5

$q = Prob(X \geq Xu)$

Table 3 Effects of double censoring on the estimator of the T-year event

Sample size N	Relative bias (%) T = 100	T = 500	T = 1000	Root mean square error (%) T = 100	T = 500	T = 1000
			p = q = 0			
30	1.4	0.04	-0.7	18.7	24.1	26.6
50	0.1	-0.9	-1.4	12.6	16.1	17.8
70	0.4	-0.1	-0.4	10.5	13.4	14.8
			p = q = 0.05			
30	-1.3	4.2	6.2	344.0	278.4	255.8
50	7.1	11.0	12.6	137.4	112.3	103.8
70	10.2	13.8	15.2	30.4	27.9	27.5
			p = q = 0.10			
30	-13.3	-0.3	0.6	583.5	487.3	440.0
50	-0.5	3.8	7.1	504.4	414.1	382.5
70	-50.8	-33.9	-27.8	328.5	270.1	249.7
			p = q = 0.15			
30	-15.3	-30.8	1.3	475.1	393.3	364.5
50	-51.1	-32.3	-25.5	368.3	335.6	288.9
70	0.9	10.6	14.2	275.0	229.6	213.6

p = q = 0 : complete (uncensored) sample

For double censoring (Table 3), while more irrigularities are observed for the relative bias, a consistent pattern is observed for the RMSE. For larger sample sizes, the RMSE becomes smaller.

It should be noted that for uncensored (complete), left censored and double samples, the value of M (number of sequences among 10,000 replications where Newton's method was convergent) is mostly more than 5,000. For right censored samples, M is alway less than 5,000.

4.2 Effects of Outliers

Typical results are shown in Table 4 for the case where a = 0.05, b = 12.5 and c = 6.0.

(i) When the outlier is included in the sample both the relative bias and RMSE are very large. With these large values, it is obvious that the outlier should not be incorporated into the sample. Doing so will introduce more bias in the estimation of the T-year event, and will also make the ML less efficient.

(ii) By removing the outlier, a much better picture is obtained: both the relative bias and RMSE reduces quite significantly.

(iii) The estimation can be further improved by censoring the outlier. The relative error and RMSE further reduce their values.

Table 4 Effect of outliers on the estimator of the T-year event

Case	N	n	Relative bias (%) T = 100	T = 500	T = 1000	Root mean square error (%) T = 100	T = 500	T = 1000
i	31	30	-71.1	-611.8	-140.7	74.5	139.2	150.5
ii	30	30	10.1	21.1	12.8	22.9	27.4	29.2
iii	31	30	-1.1	-0.8	-0.7	11.8	15.0	16.4
i	51	50	-44.1	-70.4	-83.5	46.1	74.4	88.7
ii	50	50	7.4	9.4	10.1	17.4	20.9	22.3
iii	51	50	0.1	1.2	1.5	7.7	10.0	11.0
i	71	70	-31.5	-49.2	-57.7	33.1	52.3	61.8
ii	70	70	6.4	8.4	9.2	14.9	18.5	20.0
iii	71	70	2.0	3.2	3.7	8.2	10.3	11.1

Note : Prob(X≥Xu) = Prob(W≥Wu) = 0.05

From the results so obtained, it can be said that it is best to censor the coulier, followed by removing it. Either censoring or removal must apply, otherwise, unreliable estimates will result in.

5.0 SUMMARY AND CONCLUSIONS

The effects of the type-1 cwensoring on the maximum likelihood (ML) estimator of the T-year event for the Log Pearson type-3 (LP3) distribution were investigated in this study. For this purpose the ML equations were derived and solution procedure introduced, whereby the ML estimators of the parameters a,b and c were obtained. Thus the ML estimator of the T-year event was computed and used in the investigation. Monte Carlo experiments were then employed, in which the relative bias and root mean square error in estimating the T-year event were calculated from simulated samples, repeated for a large number of sequences. The effect of an existing outlier in the sample was also analyzed in Monte Carlo experiments. It was found that:

(a) Left censoring at some level may be able to reduce the bias and even the root mean square error in estimating the T-year event.

(b) In many cases, censoring increases the root mean square error. In other words, censoring has the tendency to make the method of maximum likelihood less efficient.

(c) Exceeding large flood values, treated as outliers, should be removed from the sample, or better censored. Inclusion of the outliers will produce unreliable estimates of the T-year event, in terms of both the bias and root mean square error.

ACKNOWLEDGEMENTS

The scholarship provided by the Government of the Republic of China for the Junior author to study at the Asian Institute of Technology is gratefully acknowledged. Sincere thanks are also extended to Mrs. Nantawan for her effort in typing this paper.

REFERENCES

1. Bernado, J.M., (1976), 'Psi (Digamma) function', Applied Statistics, Vol. 26, No. 3, pp. 315-317.
2. Leese, M.N., (1973), Use of censored data in the estimation of Gumbel distribution parameters for annual maximum flood series, Water Resources Research, Vol. 9, No. 6, pp. 1534-1542.
3. Moore, R.J., (1982), Derivatives of the incomplete gamma integral, Applied Statistics, Vol. 31, No. 3, pp. 330-335.
4. Phien, H.N., (1988), A Fortran routine for the computation of gamma quantiles, Advances in Engineering Software, Vol. 10, No. 3, pp. 159-164.
5. Phien, H.N. and Hira, M.A., (1983), Log Pearson type 3 distribution : parameter estimation, Journal of Hydrology, Vol. 64, pp. 25-37.
6. Phien, H.N. and Ruksasilp, W., (1981), A review of single-site models for monthly stream flow generation, Journal of Hydrology, Vol. 52, pp. 1-12.
7. Rice, J.R., (1983), Numerical Methods, Software, and Analysis, McGraw-Hill Book Company, Singapore, pp. 239-241.