doi:10.3969/j.issn.1001-4616.2018.04.006

Estimation of Parameter for Lindley Distribution Based on Interval Data

Long Bing

(School of Mathematics and Physics,Jingchu University of Technology,Jingmen 448000,China)

Abstract:Firstly,the maximum likelihood method is used to estimate the unknown parameter in Lindley distribution under Interval data,however,the explicit expression of the parameter can not be obtained. Secondly,it is proposed that EM algorithm can be used to find out estimation of the parameter,and this method has good convergence. Finally,the simulation results show that it is feasible to use EM algorithm to estimate the unknown parameter in Lindley distribution.

Key words:Lindley distribution,interval data,EM algorithm,maximum likelihood method

CLC number:O212.1 **Document code**:A **Article ID**:1001-4616(2018)04-0029-04

基于区间数据 Lindley 分布的参数估计

龙 兵

(荆楚理工学院数理学院,湖北 荆门 448000)

[**摘要**] 首先在区间数据下用极大似然法求 Lindley 分布中未知参数的估计,然而并不能得到参数的显示表达式;其次提出用 EM 算法可以很方便地求出参数估计且该估计具有良好的收敛性;最后通过随机模拟来说明用 EM 算法求 Lindley 分布中未知参数的估计是切实可行的.

[**关键词**] Lindley 分布,区间数据,EM 算法,极大似然法

The distribution of Lindley was proposed by Lindley in 1958[1-2]. It plays an important role in the reliability of stress strength model,and many mathematical properties of Lindley distribution are more flexible than the exponential distribution. It is better to use the Lindley distribution model to fit the life data in many aspects than the exponential distribution model. So it is necessary to study the statistical properties of the distribution. At present,there are some literatures about the research of Lindley distribution,which can refer to the literatures[3-6].

In the existing research results,both the Bayesian method and the classical statistics method,their sample observation values are usually specific. However,the observed data are in some intervals in many cases,and such data are called as interval data. For example,tested samples are observed for predetermined period in the life test. Finally,the number of failure can be obtained in each time interval.

Let $0=T_0<T_1<\cdots<T_{k-1}<T_k=+\infty$,and n_j is the number of failure falling in the interval $[T_{j-1},T_j)$,$j=1,2,\cdots,k$. It is assumed that the test samples are independent and identically distributed as Lindley distribution with unknown parameter,but how to estimate the unknown parameter based on Interval data is a significant question. In recent years,many scholars have done a lot of research on this question,which have brought us a lot of research achievement,such as the literatures[7-13]. Maximum likelihood method is a better method for estimation of the parameters. For Interval data,there is no explicit solution for the maximum likelihood estimation of the parameters. Therefore,people are more concerned about how to calculate maximum likelihood estimation in practical applications. In

收稿日期:2017-03-02.

基金项目:National Natural Science Foundation of China(61374080).

通讯联系人:Long Bing,associate professor,major in applied statistics and reliability. E-mail:qh-longbing@ 163.com

this paper, EM algorithm is used to estimate the parameter of Lindley distribution based on interval data. Finally, the simulation results show that this method is feasible.

1 Maximum Likelihood Estimation of Parameter

The probability density function of Lindley distribution is illustrated as follows:

$$f(x)=\frac{\theta^2}{\theta+1}(1+x)\mathrm{e}^{-\theta x},\quad x>0. \tag{1}$$

Its distribution function is

$$F(x)=1-(1+\frac{\theta}{\theta+1}x)\mathrm{e}^{-\theta x},\quad x>0. \tag{2}$$

The parameter $\theta>0$.

Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed sample in Lindley distribution(2). The number of failure samples is n_j in the interval $[T_{j-1}, T_j)$, $j=1,2,\cdots,k$, $0=T_0<T_1<\cdots<T_{k-1}<T_k=+\infty$.

Maximum likelihood method is used to estimate the unknown parameter according to these data.

Note $p_j=P(X\in[T_{j-1},T_j))=\left(1+\frac{\theta}{\theta+1}T_{j-1}\right)\mathrm{e}^{-\theta T_{j-1}}-\left(1+\frac{\theta}{\theta+1}T_j\right)\mathrm{e}^{-\theta T_j}$.

After ignoring the constant, the likelihood function is

$$L=\prod_{j=1}^{k}p_j^{n_j},$$

$$\log L=\sum_{j=1}^{k}n_j\log p_j=\sum_{j=1}^{k}n_j\log\left[\left(1+\frac{\theta}{\theta+1}T_{j-1}\right)\mathrm{e}^{-\theta T_{j-1}}-\left(1+\frac{\theta}{\theta+1}T_j\right)\mathrm{e}^{-\theta T_j}\right],$$

$$\frac{\partial\log L}{\partial\theta}=\sum_{j=1}^{k}n_j\frac{(\theta+1)^{-2}T_{j-1}\mathrm{e}^{-\theta T_{j-1}}-(\theta+1)^{-2}T_j\mathrm{e}^{-\theta T_j}-T_{j-1}(1+T_{j-1}-\frac{1}{\theta+1}T_{j-1})\mathrm{e}^{-\theta T_{j-1}}+T_j(1+T_j-\frac{1}{\theta+1}T_j)\mathrm{e}^{-\theta T_j}}{\left(1+\frac{\theta}{\theta+1}T_{j-1}\right)\mathrm{e}^{-\theta T_{j-1}}-\left(1+\frac{\theta}{\theta+1}T_j\right)\mathrm{e}^{-\theta T_j}}.$$

Let $\frac{\partial\log L}{\partial\theta}=0$, we can obtain

$$\sum_{j=1}^{k}n_j\frac{(\theta+1)^{-2}T_{j-1}\mathrm{e}^{-\theta T_{j-1}}-(\theta+1)^{-2}T_j\mathrm{e}^{-\theta T_j}-T_{j-1}(1+T_{j-1}-\frac{1}{\theta+1}T_{j-1})\mathrm{e}^{-\theta T_{j-1}}+T_j(1+T_j-\frac{1}{\theta+1}T_j)\mathrm{e}^{-\theta T_j}}{\left(1+\frac{\theta}{\theta+1}T_{j-1}\right)\mathrm{e}^{-\theta T_{j-1}}-\left(1+\frac{\theta}{\theta+1}T_j\right)\mathrm{e}^{-\theta T_j}}=0.$$

Obviously, explicit expression of the parameter θ cannot be obtained by solving the above equation. In addition, it is also difficult to prove uniqueness of the solution in the above equation. In the following, we try to use the EM algorithm to deal with this problem.

2 Estimation of the Parameter Using EM Algorithm

Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed sample in Lindley distribution(2). They fall into the interval $[T_{j-1}, T_j)$, and the number of failure samples falling in the interval $[T_{j-1}, T_j)$ is n_j, $j=1,2,\cdots,k$, $T_0=0<T_1<\cdots<T_{k-1}<T_k=+\infty$.

Sign all the random variables for X, the observation results for Y, X_{jh} is a random variable falling into the interval $[T_{j-1}, T_j)$.

The conditional density of X_{jh} is

$$f_j(t|\theta^{(i)},Y)=\frac{\frac{(\theta^{(i)})^2}{\theta^{(i)}+1}(1+t)\mathrm{e}^{-\theta^{(i)}t}}{\int_{T_{j-1}}^{T_j}\frac{(\theta^{(i)})^2}{\theta^{(i)}+1}(1+t)\mathrm{e}^{-\theta^{(i)}t}\mathrm{d}t}=\frac{\frac{(\theta^{(i)})^2}{\theta^{(i)}+1}(1+t)\mathrm{e}^{-\theta^{(i)}t}}{\left(1+\frac{\theta^{(i)}}{\theta^{(i)}+1}T_{j-1}\right)\mathrm{e}^{-\theta^{(i)}T_{j-1}}-\left(1+\frac{\theta^{(i)}}{\theta^{(i)}+1}T_j\right)\mathrm{e}^{-\theta^{(i)}T_j}}. \tag{3}$$

E step:according to the density function of Lindley distribution,it can be obtained:

$$\log f(\theta|X)=\sum_{j=1}^{k} n_j[2\log\theta-\log(\theta+1)+\log(1+X_{jh})-\theta X_{jh}],$$

$$h(\theta|\theta^{(i)},Y)\triangleq E[\log f(\theta|X)|\theta^{(i)},Y]=2n\log\theta-n\log(\theta+1)+\sum_{j=1}^{k} n_jE[\log(1+X_{jh})]-\theta\sum_{j=1}^{k} n_jE(X_{jh})=$$

$$2n\log\theta-n\log(\theta+1)+\sum_{j=1}^{k} n_j\int_{T_{j-1}}^{T_j}\log(1+t)f_j(t|\theta^{(i)},Y)\mathrm{d}t-\theta\sum_{j=1}^{k} n_j\int_{T_{j-1}}^{T_j}tf_j(t|\theta^{(i)},Y)\mathrm{d}t.$$

M step:Finding partial derivative for $h(\theta|\theta^{(i)},Y)$ on the parameter θ.

$$\frac{\partial h}{\partial\theta}=\frac{2n}{\theta}-\frac{n}{\theta+1}-\sum_{j=1}^{k} n_j\int_{T_{j-1}}^{T_j}tf_j(t|\theta^{(i)},Y)\mathrm{d}t.$$

Let $\frac{\partial h}{\partial\theta}=0$,we can obtain

$$\frac{2n}{\theta}-\frac{n}{\theta+1}=\sum_{j=1}^{k} n_j\int_{T_{j-1}}^{T_j}tf_j(t|\theta^{(i)},Y)\mathrm{d}t.$$

Taking $\theta=\theta^{(i+1)}$ on the left side of the above equation,then obtain

$$\frac{2n}{\theta^{(i+1)}}-\frac{n}{\theta^{(i+1)}+1}=\sum_{j=1}^{k} n_j\int_{T_{j-1}}^{T_j}tf_j(t|\theta^{(i)},Y)\mathrm{d}t. \tag{4}$$

It can be proved that the left side of the equation(4)is monotone decreasing,so the solution is unique in the case of solution. Thus,the iterative process is completed from $\theta^{(i)}$ to $\theta^{(i+1)}$. Given the initial value,the parameter θ can be estimated by repeatedly using(4).

The biggest advantage of EM algorithm is simple and stable. Its main purpose is to provide a simple iterative algorithm to calculate the posterior mode. The convergence of the EM algorithm is shown in the following theorem:

Theorem 1 After each iteration,the EM algorithm improves the value of the posterior density function,that is

$$f(\theta^{(i+1)}|Y)\geqslant f(\theta^{(i)}|Y).$$

Theorem 2 (1)If $f(\theta|Y)$ has an upper bound,then $L(\theta^{(i)}|Y)$ converges to some L^{*};

(2)If $h(\theta|\varphi)$ is continuous on θ and φ,in the very general conditions on L,the convergence value θ^{*} of the estimated sequence $\theta^{(i)}$ obtained by the EM algorithm is the stable point of L.

The proof of 2 theorem is shown in literature[14].

3 Stochastic Simulation

Using random simulation to produce the interval samples of Lindley distribution(2),the specific steps are as follows:

(Ⅰ)a simple random sample following Lindley distribution with the parameter $\theta=0.4$ is generated;

(Ⅱ)Taking $T_0=0,T_1=2,T_2=3,T_3=4,T_4=5,T_5=6,T_6=8,T_7=12,T_8=+\infty$,the number of failure samples in the interval $[0,2),[2,3),[3,4),[4,5),[5,6),[6,8),[8,12),[12,+\infty)$ is $n_1,n_2,n_3,n_4,n_5,n_6,n_7,n_8$.

In this way,the sample following Lindley distribution is produced.

If a interval sample with capacity of n is generated each time,the deviation can be calculated. The initial value of the parameter is taken as $\theta^{(0)}=0.2$,the estimated values of the parameter are obtained through 4 iterations,and 4 simulation results are shown from Table 1 to Table 4 for each fixed n.

Table 1 Simulation results of n=1 000

Iteration times	First simulation		Second simulation		Third simulation		Fourth simulation	
	$\hat\theta$	deviation	$\hat\theta$	deviation	$\hat\theta$	deviation	$\hat\theta$	deviation
1	0.406 1	0.015 2	0.376 6	0.058 5	0.381 8	0.045 5	0.395 1	0.012 3
2	0.420 4	0.051 0	0.390 7	0.023 3	0.394 9	0.012 8	0.408 7	0.021 8
3	0.421 1	0.052 8	0.391 4	0.021 5	0.395 5	0.011 3	0.409 3	0.023 3
4	0.421 1	0.052 8	0.391 5	0.021 3	0.395 6	0.011 0	0.409 3	0.023 3
mean value	0.417 2	0.043 0	0.387 6	0.031 2	0.392 0	0.020 2	0.405 6	0.020 2

Table 2　Simulation results of $n=500$

Iteration times	First simulation		Second simulation		Third simulation		Fourth simulation	
	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation
1	0.399 7	0.000 8	0.364 1	0.089 8	0.361 4	0.096 5	0.373 5	0.066 3
2	0.413 6	0.034 0	0.376 8	0.058 0	0.375 4	0.061 5	0.386 4	0.034 0
3	0.414 3	0.035 8	0.377 4	0.056 5	0.376 1	0.059 8	0.387 0	0.032 5
4	0.414 3	0.035 8	0.377 4	0.056 5	0.376 2	0.059 5	0.387 0	0.032 5
mean value	0.410 5	0.026 6	0.373 9	0.065 2	0.372 3	0.069 3	0.383 5	0.041 3

Table 3　Simulation results of $n=300$

Iteration times	First simulation		Second simulation		Third simulation		Fourth simulation	
	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation
1	0.372 8	0.068 0	0.393 1	0.017 3	0.399 4	0.001 5	0.370 8	0.073 0
2	0.389 0	0.027 5	0.411 5	0.028 8	0.415 5	0.038 8	0.384 8	0.038 0
3	0.390 0	0.025 0	0.412 6	0.031 5	0.416 4	0.041 0	0.385 6	0.036 0
4	0.390 0	0.025 0	0.412 6	0.031 5	0.416 4	0.041 0	0.385 6	0.036 0
mean value	0.385 5	0.036 4	0.407 5	0.027 3	0.412 0	0.030 6	0.381 7	0.045 8

Table 4　Simulation results of $n=100$

Iteration times	First simulation		Second simulation		Third simulation		Fourth simulation	
	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation	$\hat{\theta}$	deviation
1	0.373 3	0.066 8	0.386 3	0.034 3	0.362 0	0.095 0	0.367 9	0.080 3
2	0.384 8	0.038 0	0.401 3	0.003 3	0.378 5	0.053 8	0.381 0	0.047 5
3	0.385 4	0.036 5	0.402 0	0.005 0	0.379 6	0.051 0	0.381 7	0.045 8
4	0.385 4	0.036 5	0.402 1	0.005 3	0.379 6	0.051 0	0.381 7	0.045 8
mean value	0.382 2	0.044 5	0.397 9	0.012 0	0.374 9	0.062 7	0.378 1	0.054 9

It can be seen that a satisfactory estimation of the parameter θ after 4 iterations. Whether the sample size is large or small, for the interval data that follow Lindley distribution, estimation of the parameter θ can be obtained from the above simulation process by using the iterative formula(4). And the convergence speed is faster. A better estimation after 4 iterations can be obtained, and the estimated value of the parameter is independent of the selection of the initial value.

[参考文献]

[1] LINDLEY D. Introduction to probability and statistics fron a Bayesian viewpoint, part Ⅱ: inference[M]. New York: Cambridge University Press, 1965.

[2] LINDLEY D. Fiducial distributions and Bayes' theorem[J]. Journal of the royal statistical society, 1958, 20(1): 102-107.

[3] GHITANY M E, ATIEH B, NADARAJAH S. Lindley distribution and its application[J]. Mathematics and computers in simulation, 2008, 78(4): 493-506.

[4] GHITANY M E, ALMDK, NADARAJAH S. Zero-truncated Poisson-Lindley distribution and its application[J]. Mathematics and computers in simulation, 2008, 79(3): 279-287.

[5] ZAMANI H, ISMAIL N. Negative binomial-Lindley distribution and its application[J]. Journal of mathematics and statistics, 2010, 6(1): 4-9.

[6] HUANG W P, ZHOU J L. Parameter estimation of Lindley distribution with competing risk data[J]. Systems engineering and electronic, 2016, 38(2): 464-469.

[7] JIE M, ARLENE N R. Inferences about the scale parameter of the gamma distribution based on data mixed from censoring and grouping[J]. Statistics & probability letters, 2003, 62(3): 229-243.

[8] GANG L, ZHANG C H. Linear regression with interval censored data[J]. Ann statist, 1998, 26: 1306-1327.

[9] GAETAN C, YAO J F. A multiple-imputation Metropolis version of the EM algorithm[J]. Biometrika, 2003, 90(3): 643-654.

[10] JIAN H, ROSSINI A J. Sieve estimation for the proportional-odds failure-time regression model with interval censoring[J]. JASA, 1997, 92: 960-967.

[11] RABINOWITZ D, TSIATIS A, ARAGON J. Regression with interval censored data[J]. Biometrika, 1995, 82: 501-513.

[12] ZHENG M, ZHANG H, YANG Y. Estimating parameter in some distributions from grouped data[J]. Journal of Fudan university (natural science), 2005, 44(3): 466-470.

[13] ZHENG M, XIANG Y. Estimating regression coefficients in linear models based on grouped data[J]. Mathematica applicata, 2006, 19(2): 296-303.

[14] LAWLESS J F. Statistical models and methods for lifetime data[M]. New York: Wiley, 2003.

[责任编辑:陆炳新]