第32卷第3期 上海师范大学学报(自然科学版) Vol. 32, No. 3
2003年9月 Journal of Shanghai Normal University(Natural Sciences) Sep. 2003

A New Way to Estimate the Confidence Interval for the Mean of the Exponential Distribution Based on Grouped Data

XU Hai-yan, FEI He-liang

(Mathematics and Sciences College, Shanghai Normal University, Shanghai 200234, China)

Abstract: Based on grouped data, the asymptotic normality of the maximum likelihood estimate (MLE) for mean of the single-parameter exponential distribution is proved from a new point of view, and the asymptotic confidence interval is derived. Comparing the results of CHEN and MIE, a Monte carlo simulation shows that it is a little more effective.

Key words: exponential distribution; polynomial distribution; grouped data; maximum likelihood estimate; asymptotic normality; asymptotic confidence interval

CLC number: O213.2 **Document code**: A **Article ID**: 1000-5137(2003)03-0007-06

1 Introduction and preliminaries

The exponential distribution has an important position in lifetime models. Many authors have contributed to estimate the parameters of this distribution. For Type-I and Type-II censoring, several statistical analysis methods have been developed including MLE, BLUE, BLIE and approximated MLE, see Lawless(1982)[1], Bain and Englhart(1991)[2]. For multiple Type-II censoring, Balasabramanian and Balakrishnan(1992)[4] have discussed BLUE and approximate MLE, Fei and Kong (1994)[5] have provided approximate and exact interval estimations for the exponential parameters.

In practice, it is often impossible to continuously observe or inspect the testing process even with censoring or multiple censoring. So, at this time, we might only be able to use means of grouped data. Based on grouped data, CHEN and CHENG (1988)[6] have considered the problem about Weibull parameters estimation. When it comes to estimating the approximate confidence interval for mean of the single-parameter exponential distribution, WANG and WANG (1993)[7] have transformed the grouped data into the complete data based on binomial distribution. But it is only effective for the equal-distance grouped data, and the estimation

Received date: 2003-05-21

Foundation item: Supported by the Nation Natural Science Foundation of China (69971016) and Shanghai Science and Technology Development Foundation (00JC14057) and the Special Foundation for Major Specialties of Shanghai Education Committee.

Biography: XU Hai-yan(1978-), female, doctor, Mathematics and Sciences College, Shanghai Normal University. FEI He-liang(1938-), male, professor, Mathematics and Sciences College, Shanghai Normal University.

is also very conservative. The main idea of CHEN and MEI(1996) [8] comes from the following. When they tried to search for the MLE of the mean, they needed to solve the function, such that

$$g(\theta) = \sum_{j=1}^{k} n_j \Delta\tau_j / \exp f(-\Delta\tau_j/\theta) = \sum_{j=1}^{k} n_j\tau_j + n_{\mid}k+1\mid\tau_k$$

(where $0 \equiv \tau_0 < \tau_1 < \cdots < \tau_k$ are observed times, $\Delta\tau_j = \tau_j - \tau_{j-1}$, n_j is the number of the times observed in the interval $[\tau_{j-1},\tau_j)$, and $\exp f(-\Delta\tau_j/\theta) \equiv 1 - \exp(-\Delta\tau_j/\theta)$).

They found that $\sum_{j=1}^{k} n_j\tau_j + n_{k+1}\tau_k$ isthe sum of n *i. i. d.* random variables

$$\text{s. t.} \qquad y_i = \sum_{j=1}^{k} \tau_j \Gamma(\tau_{j-1} < x_i < \tau_j) + \tau_k \Gamma(x_i > \tau_k)$$

(where $\Gamma(\cdot)$ is the identification function, $x_1,\cdots,x_n \overset{i.i.d.}{\sim} \exp(\theta)$).

Thus, according to the central limit theorem, the interval estimation of $g(\theta)$ can be got. Then the interval estimate of θ can be obtained based on the monotone of $g(\theta)$.

In this paper, we use a classical way to discuss the asymptotic normality of the MLE from a new point of view. As for CHEN[9], he has used a complete sample to prove the theorem on the asymptotic normality of MLE in Section 3, Chapter 4. We also need a complete sample to certify the conditions of the theorem. Obviously, grouped data sample is not a complete sample. But, if we change our mode of thinking, we can find that n samples with size N ($(x_{i1},\cdots,x_{in})$, $i = 1,\cdots,N$) can be looked upon as a single sample with size N ($X_1,\cdots,X_N$, $X_i = (x_{i1},\cdots,x_{in})'$, $i = 1,\cdots,N$). The distribution of the sample is polynomial distribution PN ($n,p_1,\cdots,p_{k+1}$) (where, k + 1 is the number of intervals of the grouped data) and obviously it contains only one parameter. Thus, $(X_1,\cdots,X_N) \overset{i.i.d.}{\sim} PN(n,p_1,\cdots,p_{k+1})$. Then, we can use the method of [9] to test the asymptotic normality of the MLE from the polynomial distribution. So, we can obtain the asymptotic normality of MLE from the n samples with each size N based on exponential distribution. In the later part of this paper, we prove that the MLE derived from polynomial distribution is the same as MLE directly from exponential distribution.

In Section 2, the single-parameter exponential distribution with grouped data is linked with polynomial distribution. In Section 3, MLE of the exponential parameter with grouped data is derived based on the polynomial distribution and its asymptotic normality is also proved. Thus the approximate confidence interval of the parameter is given. The results of a Monte Carlo study are given in Section 4. New confidence intervals are shown to be a little more effective than the estimators proposed in [8].

2 Distribution

Suppose that n independent observations are made on a random variable with a single-parameter exponential distribution

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{\theta}}, & x \geqslant 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.1}$$

where $\theta > 0$. Further, let $0 \equiv t_0 < t_1 < \cdots < t_k < t_{k+1} \equiv \infty$ be the ordered inspection-time and n_i be the number of failures in $(t_{i-1},t_i]$, $i = 1,2,\cdots,k+1$. Then, the likelihood function is

$$h(\theta) = \frac{n!}{n_1!n_2!\cdots n_k!n_{k+1}!} \prod_{i=1}^{k} (e^{-t_{i-1}/\theta} - e^{-t_i/\theta})^{n_i} (e^{-t_k/\theta})^{n_{k+1}} \tag{2.2}$$

That is to say the random vector $(n_1, n_2, \cdots, n_{k+1})$ has the polynomial distribution $P_N(p_1, p_2, \cdots, p_{k+1})$, where $p_i = e^{-t_{i-1}/\theta} - e^{t_i/\theta}$, $i = 1, 2, \cdots, k$ and $p_{k+1} = e^{t_k/\theta}$. Let the logarithm of $h(\theta)$ be denoted by $L(\theta)$. Then

$$L(\theta) = \log \frac{n!}{n_1! n_2! \cdots n_k! n_{k+1}!} + \sum_{i=1}^{k} n_i \log(e^{-t_{i-1}/\theta} - e^{-t_i/\theta}) - \frac{n_{k+1} t_k}{\theta}.$$

and

$$\frac{\mathrm{d}L}{\mathrm{d}\theta} = \sum_{i=1}^{k} n_i \frac{t_{i-1} e^{-t_{i-1}/\theta} - t_i e^{-t_i/\theta}}{(e^{-t_{i-1}/\theta} - e^{-t_i/\theta})\theta^2} + \frac{n_{k+1} t_k}{\theta^2}.$$

From(2.1), based on the polynomial distribution, the likelihood equation of θ is

$$H(\theta) \propto \prod_{i=1}^{N} \frac{n!1}{n_{i,1}! n_{i,2}! \cdots n_{i,k}! n_{i,k+1}!} \prod_{j=1}^{k} (e^{-t_{i-1}/\theta} - e^{-t_i/\theta})^{n_{ij}} (e^{-t_k/\theta})^{n_{i,k+1}}$$

$$\propto \prod_{i=1}^{N} (e^{-t_{i-1}/\theta} - e^{-t_i/\theta})^{\sum_{j=1}^{k} n_{ij}} (e^{-t_k/\theta})^{\sum_{j=1}^{N} n_{i,k+1}} \tag{2.3}$$

where $\sum_{j=1}^{k+1} n_{ij} = n$, $i = 1, 2, \cdots, N$, and $\sum_{j=1}^{N} n_{ij}$ represents the number of $n \times N$ products which fall into the jth interval, $j = 1, 2, \cdots, k+1$. Let $LH(\theta) = \log H(\theta)$. Then

$$\mathrm{d}LH(\theta)/\mathrm{d}\theta = \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij} \frac{t_{j-1} e^{-t_{j-1}/\theta} - t_j e^{-t_j/\theta}}{(e^{-t_{j-1}/\theta} - e^{-t_j/\theta})\theta^2} + \sum_{i=1}^{N} \frac{n_{i,k+1} t_k}{\theta^2}.$$

Let $\mathrm{d}LH(\theta)/\mathrm{d}\theta = 0$. We obtain that

$$\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij} \frac{t_j - t_{j-1}}{1 - e^{-(t_j - t_{j-1})/\theta}} = \sum_{i=1}^{N} n_{i,k+1} t_k + \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij} t_i. \tag{2.4}$$

Name the left side of equation (2.4) as $g(\theta)$. It is easy to verify that $g(\theta)$ is bigger than the right side of equation (2.4) as θ approaches to ∞. On the other hand $g(\theta)$ is less than the right side of equation (2.4) as θ approaches to 0. At the same time, $g(\theta)$ is a strictly monotone function. Hence, the solution of equation (2.4) is unique, and the solution is denoted by $\hat{\theta}_P$.

3 Asymptotic normality & asymptotic confidence interval

Lemma Equation (2.2) is the distribution of the random vector $(n_1, n_2, \cdots, n_{k+1})$.

The lemma can be proved directly from Section 2.

Theorem According to the distribution (2.2), $\hat{\theta}_P$ of θ satisfies that:

$$\sqrt{N}(\hat{\theta}_P - \theta) \to N(0, I^{-1}(\theta)),$$

where N represents N observation vectors of size n, $I(\theta)$ is Fisher information function.

Proof Since $\hat{\theta}_P$ is the unique solution of likelihood function (2.3), we only need to verify that distribution (2.2) satisfies the five conditions (see Theorem 4.6 and Theorem 4.9 in[9]) as follows.

(1) Suppose that $x_1, x_2, \cdots, x_n$ are *iid* random vectors coming from a totality whose distribution is $f(x, \theta)\mathrm{d}\mu(x)$. The parameter space θ is one-dimension open interval. Then, $f(x, \theta) > 0$ and $\mathrm{d}f(x, \theta)/\mathrm{d}\theta$ exists $(\forall \theta \in \Theta, x \in \mathscr{X})$. Furthermore, $\int_{\mathscr{X}} | \log f(x, \theta_0) | f(x, \theta_0) \mathrm{d}\mu < \infty$, where θ_0 is the true parameter.

For $E[n_i] = n(e^{-t_{i-1}/\theta} - e^{-t_i/\theta}) \leqslant n$

$$\int_{\mathscr{X}} | L(x, \theta_0) | h(x, \theta_0) \mathrm{d}\mu$$

$$= E | L(x, \theta_0) |$$

$$= E\left| \sum_{i=1}^{n} \log i - \sum_{j=1}^{k+1} \sum_{i=1}^{n} \log i \sum_{j=1}^{k} n_i \log(e^{-t_{i-1}/\theta} - e^{-t_i/\theta}) - \frac{n_{k+1} t_k}{\theta} \right|$$

$$\leqslant (k+2)\frac{(1+n)n}{2} - n\sum_{i=1}^{k} n_i (e^{-t_{i-1}/\theta} - e^{-t_i/\theta}) \log(e^{-t_{i-1}/\theta} - e^{-t_i/\theta}) + n\sum_{\theta}^{t_k} e^{-t_k/\theta} < \infty$$

And obviously $f(x,\theta) > 0$ $\mathrm{d}f(x,\theta)/\mathrm{d}(\theta)$ exists ($\forall \theta \in \Theta$, $x \in \mathscr{X}$). Therefore condition (1) is satisfied.

(2) $I(\theta) > 0$.

$$I(\theta) = -E\left(\frac{\mathrm{d}^2 L(\theta)}{\mathrm{d}\theta^2}\right)$$

$$= E\left\{\frac{2}{\theta^3}\left[-\sum_{i=1}^{k} n_i \frac{t_i - t_{i-1}}{1 - e^{-(t_i - t_{i-1})/\theta}} + n_{k+1} t_k + \sum_{i=1}^{k} n_i t_i\right] + \frac{1}{\theta^4}\sum_{i=1}^{k} n_i \frac{(t_i - t_{i-1})^2 e^{-(t_i - t_{i-1})/\theta}}{(1 - e^{-(t_i - t_{i-1})/\theta})^2}\right\}$$

$$= \frac{n}{\theta^4}\sum_{i=1}^{k} \frac{(t_i - t_{i-1})^2}{e^{t_i/\theta} - e^{t_{i-1}/\theta}} > 0$$

(3) $E\dfrac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} = 0$

$$E\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}$$

$$= E\left(\sum_{i=1}^{k} n_i \frac{t_{i-1} e^{-t_{i-1}/\theta} - t_i e^{-t_i/\theta}}{(e^{-t_{i-1}/\theta} - e^{-t_i/\theta})\theta^2} + \frac{n_{k+1} t_k}{\theta^2}\right)$$

$$= \frac{n}{\theta^2}\left[\sum_{i=1}^{k} (t_{i-1} e^{-t_{i-1}/\theta} - t_i e^{-t_i/\theta}) + t_k e^{-t_k/\theta}\right]$$

$$= 0.$$

(4) $\dfrac{\mathrm{d}^2 L(\theta)}{\mathrm{d}\theta^2}$ is obviously continuous based on $\theta(\theta \in \Theta)$.

(5) $\forall\ \varepsilon$ (may be the function of θ), such that $E_\theta[J(x,\theta,\varepsilon] < \infty$, where $J(x,\theta,\varphi) = \sup\{|\partial^2 \log f(x,\varphi)/\partial\varphi^2| : |\varphi - \theta| < \varepsilon\}$.

For

$$|\mathrm{d}^2 L(\varphi)/\mathrm{d}\varphi^2| = \left| -\frac{2}{\varphi^3} n_{k+1} t_k + \sum_{i=1}^{k} n_i \left[\frac{t_{i-1}^2 e^{-t_{i-1}/\varphi} - t_i^2 e^{-t_i/\varphi}}{\varphi^4 (e^{-t_{i-1}/\varphi} - e^{-t_i/\varphi})} - \frac{(t_{i-1} e^{-t_{i-1}/\varphi} - t_i e^{-t_i/\varphi})^2}{\varphi^4 (e^{-t_{i-1}/\varphi} - e^{-t_i/\varphi})^2} - \frac{2}{\varphi^3}\frac{t_{i-1} e^{-t_{i-1}/\varphi} - t_i e^{-t_i/\varphi}}{e^{-t_{i-1}/\varphi} - e^{-t_i/\varphi}}\right]\right|$$

$$= \left| \sum_{i=1}^{k} n_i \frac{1}{2}\eta_i^2 - 2\eta_i\varphi - \{(\xi_i - \varphi)\}^2 - 2\varphi\{(\xi_i - \varphi)\}\}\{\varphi^4\} - \frac{2}{\varphi^3} n_{k+1} t_k \right|$$

(Based on the Theorem of Mean. Let $\eta_i, \xi_i \in [t_{i-1}, t_i)$)

$$= \left| \sum_{i=1}^{k} n_i \frac{-\xi_i^2 + \eta_i^2 - 2\eta_i\varphi + \varphi^2}{\varphi^4} - \frac{2}{\varphi_3} n_{k+1} t_k \right|$$

$$\leqslant \sum_{i=1}^{k} n_i (2t_i^2/\varphi^4 + 2t_i/\varphi^3 + 1/\varphi^2) + \frac{2}{\varphi^3} n_{k+1} t_k$$

$$\leqslant \sum_{i=1}^{k} n_i (32t_i^2/\theta^4 + 16t_i/\theta^3 + 4/\theta^2) + 16\frac{2}{\theta^3} n_{k+1} t_k, \quad \varepsilon = \theta/2$$

Hence $E_\theta J(x,\theta,\varphi) < n\sum_{i=1}^{k} (32t_i^2/\theta^4 + 16t_i/\theta^3 + 4/\theta^2) + 16n\dfrac{t_k}{\theta^3} < \infty$.

Correspondingly, Theorem 2 is proved.

Corollary The MLE of θ which comes from distribution (2.1) based on a sample of size $n \times N$ with

grouped data is $\hat{\theta}_N$. And, it has the same asymptotic normality as $\hat{\theta}_P$.

Proof According to equation (2.3), we can get that the likelihood function of a sample with size $N(X_1, \cdots, X_N)$ based on distribution (2.2) is identical to the likelihood function of a sample with size $n \times N$ based on distribution (2.1) with grouped data. Thus, the MLE of θ which comes from distribution (2.2) is identical to the MLE of θ which comes from the distribution (2.1) based on a sample of size $n \times N$ with grouped data. At the same time, the fisher information matrices are the same. Thus, the corollary is verified.

Replace θ by $\hat{\theta}_P$. The $1-\alpha$ approximate confidence interval of θ is obtained:

$$(\hat{\theta}_P - \frac{u_{\alpha/2}}{\sqrt{NI(\hat{\theta}_P)}},\quad \hat{\theta}_P + \frac{u_{\alpha/2}}{\sqrt{NI(\hat{\theta}_P)}}),$$

where $u_{\alpha/2}$ is the $\alpha/2$ - up quartile of the standard normal distribution.

4 Monte Carlo Simulation and Comparison

To compare the estimate proposed in this article (based on the polynomial distribution) with other commonly used estimates, a Monte Carlo study was undertaken and several cases were considered: When $\theta = 1$, and $\theta = 2$, the inspection-time sequence was $t = (0.1\ 0.3\ 0.5\ 0.7\ 1.0\ 1.3\ 1.5\ 1.6\ 1.8\ 2.0)$, when $\theta = 5$, and $\theta = 20$, $t = (1\ 2\ 4\ 5\ 7\ 10\ 12\ 15\ 17\ 20)$, and when $\theta = 10$, $t = (1\ 2.5\ 5\ 7.5\ 10\ 12\ 14\ 20\ 21\ 24\ 30)$.

In Tables 1 ~ 2, for different size of sample and different θ's, we compare θ's MLE based on the polynomial distribution (denoted by $\hat{\theta}_P$) and directly from the exponential distribution (denoted by $\hat{\theta}_E$). At the same time, we compare the 95% confidence interval of θ relied on the method proposed by us (denoted by N-interval) and the method from CHENG and MIE[6] (denoted by C- interval). The meaning of n, N and M are the same as what are proposed in Section 3.

Table 1　$n \times N \times M = 10 \times 30 \times 30$

True Value	$\theta = 1$	$\theta = 5$	$\theta = 10$	$\theta = 20$
$\hat{\theta}_P$	1.0026	5.0114	10.1133	19.8784
$\hat{\theta}_E$	1.0125	4.979	10.0384	20.2879
$(\hat{\theta}_P - \theta)^2$	0.0034	0.1122	0.3375	1.6653
$(\hat{\theta}_E - \theta)^2$	0.0043	0.0697	0.4743	2.0707
N-interval	(0.8874,1.1345)	(4.4817,5.6460)	(8.5949,10.8599)	(17.1810,22.8990)
C-interval	(0.8992,1.1651)	(4.3631,5.5005)	(8.7629,11.1564)	(17.1801,23.6773)

Table 2　$n \times N \times M = 1 \times 50 \times 30$

True Value	$\theta = 1$	$\theta = 5$	$\theta = 10$	$\theta = 20$
$\hat{\theta}_P$	1.0013	2.0595	4.87	20.7228
$\hat{\theta}_E$	0.9652	1.9320	5.2705	20.4201
$(\hat{\theta}_P - \theta)^2$	0.0224	0.1016	0.3643	9.4865
$(\hat{\theta}_E - \theta)^2$	0.0147	0.1411	0.6279	8.4380
N-interval	(0.5853,1.3184)	(1.0832,2.8859)	(3.5518,63385)	(13.5981,28.6588)
C-interval	(0.7649,1.4665)	(1.4540,3.2251)	(3.4897,6.3152)	(14.6666,33.3356)

By Monte Carlo simulation we conclude that:

(1) MLE from the polynomial distribution approximates to that comes directly from the exponential distribution, which verifies the corollary.

(2) It is obvious that approximate confidence interval of θ from the polynomial distribution is a little bit prior to [6], especially when the true value of θ is so different from the mean of the inspection vector t.

Reference:

[1] LAWLESS J F. Statistical Models and Method for Lifetime Data[M]. John Wiley & Sons, 1982.

[2] BAIN L JM, ENGELJARDT M. Statistical Analysis of Reliability and Life-Testing Models-Theory and Methods[M]. Marce l Dekker, 1991.

[3] BALAKRISHNAN N. On the Maximum Likelihood Estimation of the Location and Scale Parameters Based on Multiply Type-II Censoring Samples[J]. J Appl Statist, 1990, 17:55-61.

[4] BALASUBRAMANIAN K, BALAKRISHNAN N. Estimate for One and Two-Parameter Exponential Distribute Under Multi - ple Type-II Censoring[J]. Statistical Papers, 1992, 33:203-216.

[5] FEI H L, KONG F. Interval Estimations for One and Two-Parameter Exponential Distributions Under Multiple Type-II Censoring[J]. Commun Statist -Theory Meth, 1994, 23:1717-1733.

[6] CHENG K F,CHEN C H. Estimation of the Weibull Parameter with Grouped Data[J]. Commun Statist Theory Meth, 1994, 17:325-341.

[7] WANG B X, WANG L L. Confidence Intervals for Mean Life Based on Grouped Data under Exponential Distribution[J]. Chinese Journal of Applied Probability and Statistics, 1988:415-422.

[8] CHEN Z M, MIE J. Confidence Interval for the Mean of the Exponential Distribution Based on Grouped Data[J]. IEEE Transactions on Reliability, 1996,45:671-677.

[9] CHEN X R. Advanced Mathematical Statistics[M]. Publishing House of University of Science & Technology of China, 1999.

指数分布场合下基于分组数据区间估计的新方法

徐海燕，费鹤良

(上海师范大学 数理信息学院，上海 200234)

摘要：从新的角度证明了分组数据下指数分布总体均值的极大似然估计(MLE)的渐进正态性，给出了该均值的渐进置信区间. 通过 Monte Carlo 模拟比较，发现该置信区间优于 CHEN 和 MIE 得到的置信区间.

关键词：指数分布；多项分布；分组数据；极大似然估计；渐进正态性；渐进置信区间