Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2009, Article ID 310575, 16 pages
doi:10.1155/2009/310575

Research Article

On the Existence and Uniqueness of the Maximum Likelihood Estimators of Normal and Lognormal Population Parameters with Grouped Data

Jin Xia,[1] Jie Mi,[2] and YanYan Zhou[3]

[1] *Center on Aging and Health, John Hopkins University, Baltimore, MD 21287, USA*
[2] *Department of Statistics, Florida International University, Miami, FL 33139, USA*
[3] *Department of Statistics & Biostatistics, California State University, Hayward, CA 94542, USA*

Correspondence should be addressed to YanYan Zhou, yanyan.zhou@csueastbay.edu

Received 11 March 2009; Accepted 16 June 2009

Recommended by A. Thavaneswaran

Lognormal distribution has abundant applications in various fields. In literature, most inferences on the two parameters of the lognormal distribution are based on Type-I censored sample data. However, exact measurements are not always attainable especially when the observation is below or above the detection limits, and only the numbers of measurements falling into predetermined intervals can be recorded instead. This is the so-called grouped data. In this paper, we will show the existence and uniqueness of the maximum likelihood estimators of the two parameters of the underlying lognormal distribution with Type-I censored data and grouped data. The proof was first established under the case of normal distribution and extended to the lognormal distribution through invariance property. The results are applied to estimate the median and mean of the lognormal population.

1. Introduction

Lognormal distribution has been used to model many skewed frequency distributions, especially to model continuous random quantities in medical, physical, chemical, biological, toxicological, economical, and environmental processes.

For example, in medicine, the red cell volume distributions; size distributions of plaques in Alzheimer's patients; surgical procedure times; survival times of breast and ovarian cancer; all have been modeled by lognormal distribution by various researchers. Tai et al. [1] and Mould et al. [2] validated the use of the lognormal model for predicting long-term survival rates of laryngeal cancer patients using short-term follow-up data.

It is also common to apply the lognormal distribution for fatigue life and residual strength of composite materials [3], reliability analysis [4], size distributions in economics and actuarial sciences [5], cell growth [6], and many other phenomena.

In all these studies, it is critical to estimate the parameters of a lognormal distribution. A random variable follows lognormal distribution $LN(\mu,\sigma)$ if the logarithm of the random variable follows normal distribution $N(\mu,\sigma)$. Thus to estimate the parameters (μ,σ), it suffices to convert the lognormal data to normal data by log-transformation. In literature, the estimation of these two parameters was considered with complete sample, or in most cases Type-I censored sample. However, estimation with grouped data has not yet been studied. We complement this literature by proposing maximum likelihood estimators (MLEs) of the two parameters that are based on grouped sample data (i.e., interval censored data).

The paper is organized as follows. In Section 2, we will show that the MLEs of the two parameters exist uniquely under mild conditions and thus the asymptotic normality of the estimators. The results are applied to derive the point and confidence interval estimation of the mean and median of the underlying lognormal distribution in Section 2.1. Section 3 provides the simulation results comparing the properties of the estimator based on grouped sample to those of type I censoring. Section 4 contains study results of a practical problem by the above method. To facilitate reading, proofs are relegated to the appendix.

2. Main Results

In this section, we will first show that the MLEs of the parameters μ and σ of a normal population $N(\mu,\sigma^2)$ based on grouped data uniquely exist. Here, the grouped data refers to the following. Assume that a sample $X_1,\ldots,X_n$ is drawn from a normal population, the values of X_js are unknown; however, according to k preestablished partition points $\tau_1 < \tau_2 < \cdots < \tau_k$, we know n_i, the number of X_js that fall into the interval $[\tau_{i-1},\tau_i)$, $1 \le i \le k+1$ where $\tau_0 \equiv -\infty$ and $\tau_{k+1} \equiv \infty$. Denote the density of the standard normal distribution $N(0,1)$ as $\varphi(t)$, then the density of $N(\mu,\sigma^2)$ distribution is $f(t;\mu,\sigma) = (1/\sigma)\varphi((t-\mu)/\sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$. In order to prove our results, we consider two new parameters $\theta_1 = \mu/\sigma$ and $\theta_2 = 1/\sigma$. There is a one-to-one correspondence between (μ,σ) and (θ_1,θ_2), namely, $\mu = \theta_1/\theta_2$ and $\sigma = 1/\theta_2$. We will show that the MLEs of θ_1 and θ_2 based on grouped data uniquely exist. Then due to the invariance property of MLEs, the existence and uniqueness of the MLEs of (μ,σ) follow. With the new parameters (θ_1,θ_2), the CDF of $N(\mu,\sigma^2)$ can be expressed as $\Phi(\theta_2 t - \theta_1)$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution, and the log-likelihood function $\ln L$ is given by

$$\begin{aligned}\ln L = c &+ n_1 \ln \Phi(\theta_2\tau_1 - \theta_1) + n_{k+1}\ln[1-\Phi(\theta_2\tau_k-\theta_1)] \\ &+ \sum_{i=2}^{k} n_i \ln[\Phi(\theta_2\tau_i-\theta_1) - \Phi(\theta_2\tau_{i-1}-\theta_1)],\end{aligned} \tag{2.1}$$

where c is a known constant.

Before proceed, we present two lemmas. Please refer to the appendix for the proofs of the lemmas.

Lemma 2.1. *Assume $n_1 + n_{k+1} < n$, $n_{j-1} + n_j < n$, $2 \le j \le k+1$. For any given $\eta > 0$, there exists a compact subset $K \equiv K(\eta) \subset (-\infty, \infty) \times (0, \infty)$ such that*

$$\{(\theta_1, \theta_2) : \ln L(\theta_1, \theta_2) \ge -\eta\} \subset K. \tag{2.2}$$

Basically, Lemma 2.1 means that the log-likelihood function $\ln L(\theta_1, \theta_2)$ will not achieve its maximum value at the boundary of its domain.

Lemma 2.2. *Let $g(u, v) \equiv \ln(\Phi(u) - \Phi(v))$ for $v < u$. Then the Hessian matrix H^* of $g(u, v)$,*

$$H^* = \begin{pmatrix} \dfrac{\partial^2 g}{\partial u^2} & \dfrac{\partial^2 g}{\partial u \partial v} \\ \dfrac{\partial^2 g}{\partial u \partial v} & \dfrac{\partial^2 g}{\partial v^2} \end{pmatrix}, \tag{2.3}$$

is negative definite.

Theorem 2.3. *Suppose that the observed $n_1, \ldots, n_{k+1}$ satisfy $n_1 + n_{k+1} < n$ and $n_{j-1} + n_j < n$, $\forall 2 \le j \le k+1$, then the MLEs of parameters μ and σ of normal population $N(\mu, \sigma^2)$ uniquely exist.*

Proof. We need only to show that the MLEs of parameters θ_1 and θ_2 uniquely exist. According to the results of Mäkeläinen et al. [7], in order to show the existence and uniqueness of the MLEs of (θ_1, θ_2), it is sufficient to verify the following two conditions.

(i) For any given $\eta > 0$, (2.2) holds.

(ii) The Hessian matrix of $\ln L$,

$$H(\theta_1, \theta_2) = \left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right), \tag{2.4}$$

is negative definite at every point $(\theta_1, \theta_2) \in (-\infty, \infty) \times (0, \infty)$.

Condition (i) is certainly satisfied by Lemma 2.1. Therefore, to prove the theorem, we need only to show (ii), that is, the log-likelihood function $\ln L$ is negative definite function of

$$\boldsymbol{\theta} = (\theta_1, \theta_2) \in (-\infty, \infty) \times (0, \infty). \tag{2.5}$$

To this end we should consider each of the three terms in the expression (2.1) of $\ln L(\boldsymbol{\theta})$.

Let $g_1(\boldsymbol{\theta}) \equiv \ln \Phi(\theta_2\tau_1 - \theta_1)$. It is evident that the Hessian matrix of g_1 is

$$H_1 \equiv \begin{pmatrix} \dfrac{\partial^2 g_1}{\partial \theta_1^2} & \dfrac{\partial^2 g_1}{\partial \theta_1 \partial \theta_2} \\ \dfrac{\partial^2 g_1}{\partial \theta_1 \partial \theta_2} & \dfrac{\partial^2 g_1}{\partial \theta_2^2} \end{pmatrix}$$
$$= \begin{pmatrix} \dfrac{\varphi'(\theta_2\tau_1 - \theta_1)\Phi(\theta_2\tau_1 - \theta_1) - Q}{\Phi(\theta_2\tau_1 - \theta_1)^2} & -\tau_1 \dfrac{\varphi'(\theta_2\tau_1 - \theta_1)\Phi(\theta_2\tau_1 - \theta_1) - Q}{\Phi(\theta_2\tau_1 - \theta_1)^2} \\ -\tau_1 \dfrac{\varphi'(\theta_2\tau_1 - \theta_1)\Phi(\theta_2\tau_1 - \theta_1) - Q}{\Phi(\theta_2\tau_1 - \theta_1)^2} & \tau_1^2 \dfrac{\varphi'(\theta_2\tau_1 - \theta_1)\Phi(\theta_2\tau_1 - \theta_1) - Q}{\Phi(\theta_2\tau_1 - \theta_1)^2} \end{pmatrix}, \tag{2.6}$$

where Q denotes $\varphi^2(\theta_2\tau_1 - \theta_1)$.

To show H_1 is negative semidefinite, we will verify the following two conditions: (a) $\partial^2 g/\partial\theta_1^2 < 0$ or $\partial^2 g/\partial\theta_2^2 < 0$, $\forall(\theta_1, \theta_2) \in (-\infty, \infty) \times (0, \infty)$; (b) the determinant of H_1 is nonnegative, that is, $|H_1| \geq 0$.

Note that (a) is equivalent to $-(\theta_2\tau_1 - \theta_1)\Phi(\theta_2\tau_1 - \theta_1) - \varphi(\theta_2\tau_1 - \theta_1) < 0$, $\forall(\theta_1, \theta_2) \in (-\infty, \infty) \times (0, \infty)$. This is true since $y[1 - \Phi(y)] < \varphi(y)$ holds for any y (see, e.g., Feller [8]). Hence (a) is satisfied. The two rows of H_1 are proportional, so $|H_1| = 0$. Hence, the condition (b) is satisfied. Therefore, H_1 is negative semidefinite.

Now denote $g_{k+1}(\boldsymbol{\theta}) \equiv \ln[1 - \Phi(\theta_2\tau_k - \theta_1)]$. The Hessian matrix H_{k+1} of g_{k+1} is

$$H_{k+1} \equiv \begin{pmatrix} -\dfrac{\varphi'(\theta_2\tau_k - \theta_1)[1 - \Phi(\theta_2\tau_k - \theta_1)] + \varphi^2(\theta_2\tau_k - \theta_1)}{[1 - \Phi(\theta_2\tau_1 - \theta_1)]^2} & \tau_k \dfrac{\varphi'(\theta_2\tau_k - \theta_1)\mathcal{F}}{[1 - \Phi(\theta_2\tau_k - \theta_1)]^2} \\ \tau_k \dfrac{\varphi'(\theta_2\tau_k - \theta_1)\mathcal{F}}{[1 - \Phi(\theta_2\tau_k - \theta_1)]^2} & -\tau_k^2 \dfrac{\varphi'(\theta_2\tau_k - \theta_1)\mathcal{F}}{[1 - \Phi(\theta_2\tau_k - \theta_1)]^2} \end{pmatrix}, \tag{2.7}$$

where $\mathcal{F}$ denotes $[1 - \Phi(\theta_2\tau_k - \theta_1)]^2 + \varphi^2(\theta_2\tau_k - \theta_1)$.

In the similar way as the above we can show that the matrix H_{k+1} is negative semidefinite.

Finally, let us consider $h(\theta_1, \theta_2) \equiv \ln[\Phi(\theta_2\tau_i - \theta_1) - \Phi(\theta_2\tau_{i-1} - \theta_1)]$, $2 \leq i \leq k$. Let $u = \theta_2\tau_i - \theta_1$, $v = \theta_2\tau_{i-1} - \theta_1$. Then $h(\theta_1, \theta_2) = \ln[\Phi(u) - \Phi(v)] \equiv g(u, v)$. The Hessian matrix H_i associated with $h(\theta_1, \theta_2)$ is

$$H_i = \begin{pmatrix} \dfrac{\partial^2 h}{\partial \theta_1^2} & \dfrac{\partial^2 h}{\partial \theta_1 \partial \theta_2} \\ \dfrac{\partial^2 h}{\partial \theta_1 \partial \theta_2} & \dfrac{\partial^2 h}{\partial \theta_2^2} \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ \tau_i & \tau_{i-1} \end{pmatrix} \begin{pmatrix} \dfrac{\partial^2 g}{\partial u^2} & \dfrac{\partial^2 g}{\partial u \partial v} \\ \dfrac{\partial^2 g}{\partial u \partial v} & \dfrac{\partial^2 g}{\partial v^2} \end{pmatrix} \begin{pmatrix} -1 & \tau_i \\ -1 & \tau_{i-1} \end{pmatrix} = A' H^* A, \tag{2.8}$$

where

$$H^* = \begin{pmatrix} \dfrac{\partial^2 g}{\partial u^2} & \dfrac{\partial^2 g}{\partial u \partial v} \\ \dfrac{\partial^2 g}{\partial u \partial v} & \dfrac{\partial^2 g}{\partial v^2} \end{pmatrix}, \qquad A = \begin{pmatrix} -1 & \tau_i \\ -1 & \tau_{i-1} \end{pmatrix}, \tag{2.9}$$

and A' is the transpose of A. By Lemma 2.2, H^* is negative definite. Therefore, H_i is negative definite.

The Hessian matrix H of the log-likelihood function $\ln L(\boldsymbol{\theta})$ can be expressed as $H = n_1 H_1 + n_{k+1} H_{k+1} + \sum_{i=2}^{k} n_i H_i$. Since matrices H_1 and H_{k+1} are negative semidefinite, each H_i $(2 \le i \le k)$ is negative definite, and at least one $n_i > 0$ by our assumptions, so H must be negative definite. This completes the proof of the theorem. □

Corollary 2.4. *Under the conditions of Theorem 2.3, it holds that as* $n \to \infty$,

$$\begin{pmatrix} \widehat{\mu}_n \\ \widehat{\sigma}_n \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \xrightarrow{L} N\Big(\mathbf{0}, I^{-1}(\mu, \sigma)\Big), \tag{2.10}$$

where $\xrightarrow{L}$ *means "converges in law," and*

$$I(\mu, \sigma) = -E \begin{pmatrix} \dfrac{\partial^2 \ln L}{\partial \mu^2} & \dfrac{\partial^2 \ln L}{\partial \mu \partial \sigma} \\ \dfrac{\partial^2 \ln L}{\partial \mu \partial \sigma} & \dfrac{\partial^2 \ln L}{\partial \sigma^2} \end{pmatrix}. \tag{2.11}$$

Proof. For each $n > k$, define $A_n \equiv \{n_1 + n_{k+1} = n, n_{i-1} + n_i = n,\ 1 \le i \le k\}$. Note that $P(\limsup_{n\to\infty} A_n) = 0$. Hence the result follows from Theorem 2.3 and the asymptotic normality of MLE (see, e.g., Lawless (2003)).

The same results as in Theorem 2.3 and Corollary 2.4 also hold for the case of Type-I censored data. Let $X_1, \ldots, X_n$ be a sample from an $N(\mu, \sigma^2)$ population. Suppose that τ is a predetermined detection limit. Without loss of generality, we will consider left censoring, the common situation in environmental studies, that is, X_j will be observed if and only if $X_j \ge \tau$. Even though Type-I is widely applied in literature, but according to the authors' knowledge, the existence and uniqueness of the MLEs of (μ, σ) have not been proved. This will be shown in the following theorem. □

Theorem 2.5. *Suppose that the number of observable* X_j*s is at least 2, then the MLEs of* (μ, σ) *uniquely exist based on the Type-I censored data with* τ *as detection limit.*

Proof. The result can be proved in the same way as Balakrishnan and Mi [9]. □

Remark 2.6. (a) The same result holds for the case of right censoring; (b) the results of Theorem 2.5 are true if each X_j is censored by detection limit τ_j $(1 \le j \le n)$.

2.1. Estimation of the Median and Mean

Suppose that random variable Y follows lognormal distribution $LN(\mu, \sigma^2)$. With log-transformation then $X = \ln Y$ follows normal distribution $N(\mu, \sigma^2)$. Lognormal distribution has been used to model various continuous random variables as mentioned in Section 1. Specifically, this distribution is frequently applied in environmental statistics. The lognormal random variable Y has median $m \equiv \exp\{\mu\}$ and mean $\nu \equiv E(Y) = \exp\{\mu + \sigma^2/2\}$. The MLEs of m and ν can easily be obtained as $\widehat{m} = \exp\{\widehat{\mu}\}$ and $\widehat{\nu} = \exp\{\widehat{\mu} + \widehat{\sigma}^2/2\}$ due to the invariance property of MLE. We can also obtain approximate confidence intervals for m and ν as follows.

Denote the inverse of the matrix $I(\mu, \sigma)$ in Corollary 2.4 to Theorem 2.3 as

$$I^{-1}(\mu, \sigma) = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}. \tag{2.12}$$

It is obvious that $\widehat{\mu}_n - \mu \xrightarrow{L} N(0, \beta_{11})$, by large sample theory we have $\exp\{\widehat{\mu}_n\} - \exp\{\mu\} \xrightarrow{L} N(0, (\exp\{\mu\}^2)\beta_{11})$. From these, an approximate $(1-\alpha)100\%$ confidence interval of m can be obtained as $\exp\{\widehat{\mu}_n\} \pm z_{\alpha/2}(\exp\{\widehat{\mu}_n\}\sqrt{\widehat{\beta}_{11}})$, here $z_{\alpha/2}$ is the upper $\alpha/2$ percentile, and $\widehat{\beta}_{11}$ is obtained from substituting $\widehat{\mu}_n$ and $\widehat{\sigma}$ for μ and σ in the expression of β_{11}. Similarly, it holds that as $n \to \infty$

$$e^{\widehat{\mu}_n + \widehat{\sigma}_n^2/2} - e^{\mu + \sigma^2/2} \xrightarrow{L} N\left(\mathbf{0}, \tau^2\right), \tag{2.13}$$

where

$$\tau^2 = \left(e^{\mu+\sigma^2/2}, \sigma e^{\mu+\sigma^2/2}\right) I^{-1}(\mu, \sigma) \begin{pmatrix} e^{\mu+\sigma^2/2} \\ \sigma e^{\mu+\sigma^2/2} \end{pmatrix}. \tag{2.14}$$

Therefore, an approximate $(1-\alpha)100\%$ confidence interval of $\nu = \exp\{\mu + \sigma^2/2\}$ is obtained as

$$e^{\widehat{\mu}_n + \widehat{\sigma}_n^2/2} \pm z_{\alpha/2}\widehat{\tau}, \tag{2.15}$$

where $\widehat{\tau}$ is obtained by substituting μ and σ by their MLEs $\widehat{\mu}_n$ and $\widehat{\sigma}_n$.

3. Simulation Studies

In this section, we will conduct simulation studies on the MLEs and confidence intervals of μ and σ of normal distribution $N(\mu, \sigma^2)$ based on grouped data. In addition, we will also examine point and interval estimations of the mean and median of lognormal distribution $LN(\mu, \sigma^2)$. The results obtained from grouped data will be compared with those obtained from Type-I censored data.

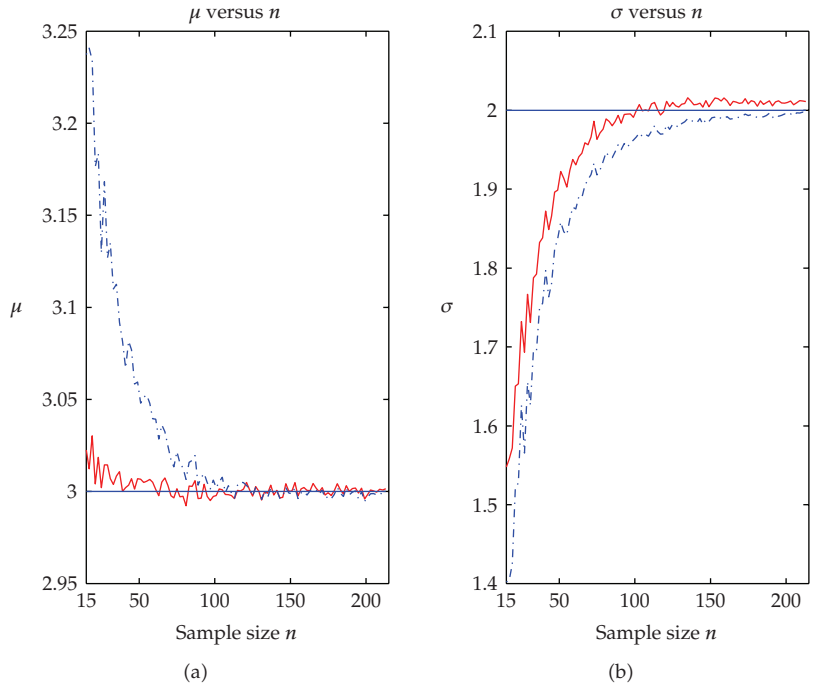

Figure 1: Grouped data: $\tau_1 = 2$, $\tau_2 = 2.5$, $\tau_3 = 3$, $\tau_4 = 3.5$, $\tau_5 = 4$. Type I left censored data: $\tau = 2$.

We create a population of size n by drawing n values from a normal population with $\mu = 3$ and $\sigma = 2$. Next, for a prefixed five partition points τ_i, $1 \le i \le 5$, we record the number of this population that fall into each interval $[\tau_{i-1}, \tau_i)$. Each such samples are consider to be observed sample. The MLEs of μ and σ are then computed based on this observed sample. This process is repeated 5,000 times. Different sample size and 6 sets of partition points are considered for comparisons purpose.

We compute the MLEs of $\theta_1 = \mu/\sigma$ and $\theta_2 = 1/\sigma$ by solving the likelihood equations

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta_1} &= 0, \\ \frac{\partial \ln L}{\partial \theta_2} &= 0, \end{aligned} \tag{3.1}$$

using SAS IMSL nonlinear equation solver. Then the MLEs $\widehat{\mu} = \widehat{\mu}_n$ and $\widehat{\sigma} = \widehat{\sigma}_n$ of μ and σ are readily obtained by the invariance of MLE. According to the large sample properties of MLEs stated in Corollary 2.4 to Theorem 2.3, we know that $(\widehat{\mu}_n, \widehat{\sigma}_n)$ is asymptotically normally distributed. Thus we can obtain approximate confidence intervals for μ and σ.

Type-I censored data are very common in various experiments. It is widely used in life test in order to save test time. Particularly, in environmental data analysis, values are often reported simply as being below detection limit along with the stated detection limit. The data obtained in this way are Type-I left singly censored. To compare the performance of the MLEs based on the grouped data with those obtained from Type-I left singly censored data, we will use τ_1 as the "detection limit". Figures 1, 2, 3, 4, 5, and 6 present the estimated MLEs of μ

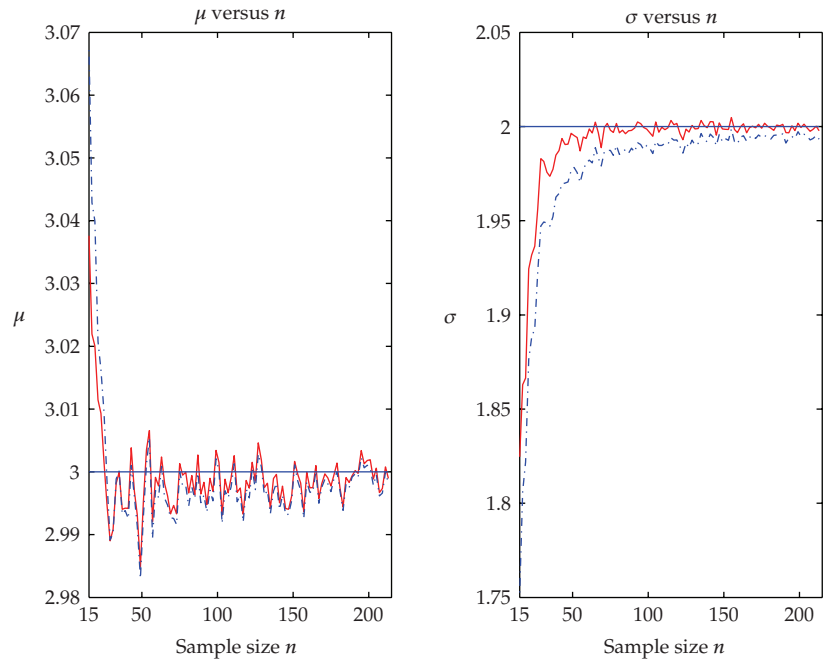

Figure 2: Grouped data: $\tau_1 = 1.5$, $\tau_2 = 2.5$, $\tau_3 = 3.5$, $\tau_4 = 4.5$, $\tau_5 = 5.5$. Type I left censored data: $\tau = 1.5$.

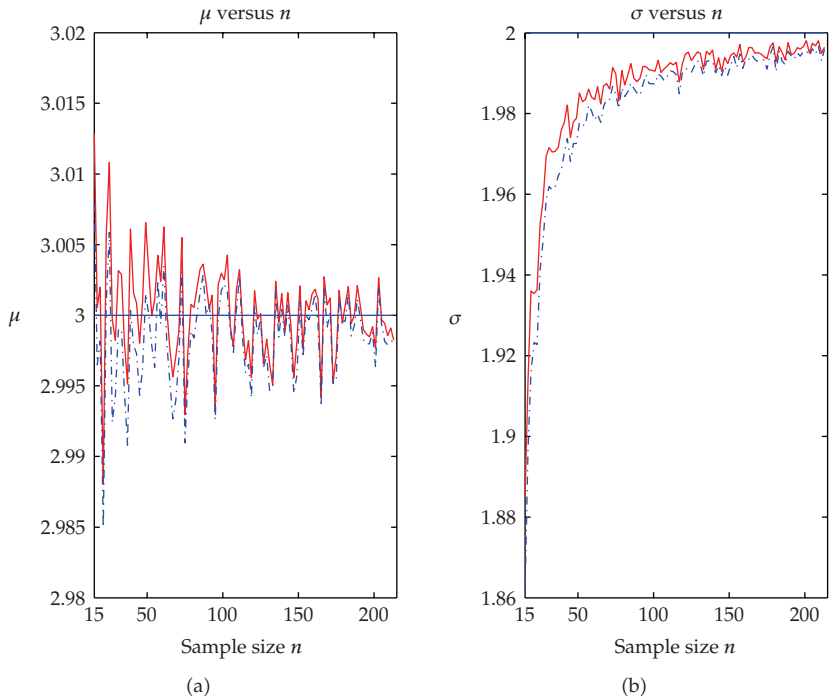

Figure 3: Grouped data: $\tau_1 = 0$, $\tau_2 = 1.5$, $\tau_3 = 3$, $\tau_4 = 4.5$, $\tau_5 = 6$. Type I left censored data: $\tau = 0$.

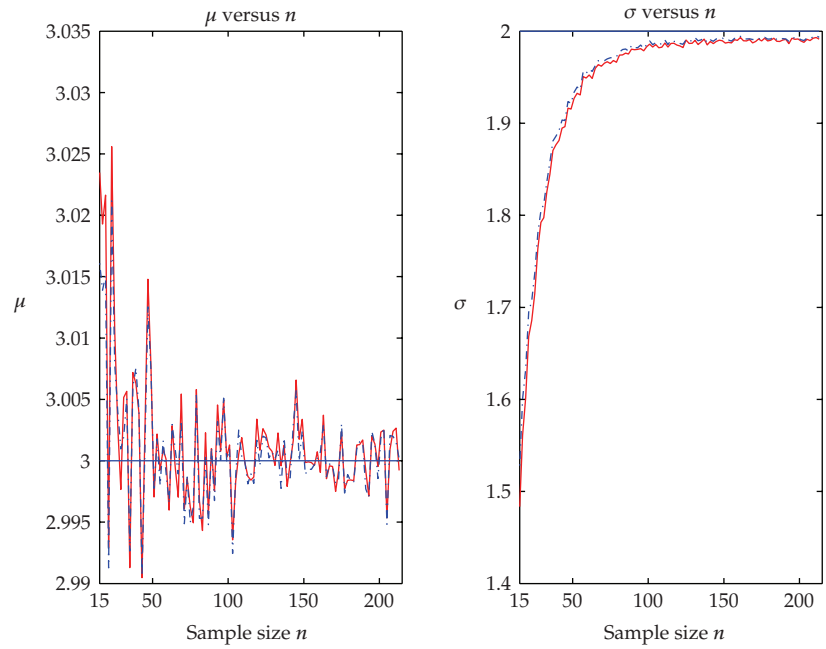

Figure 4: Grouped data: $\tau_1 = -1.5$, $\tau_2 = 1$, $\tau_3 = 3.5$, $\tau_4 = 6$, $\tau_5 = 8.5$. Type I left censored data: $\tau = -1.5$.

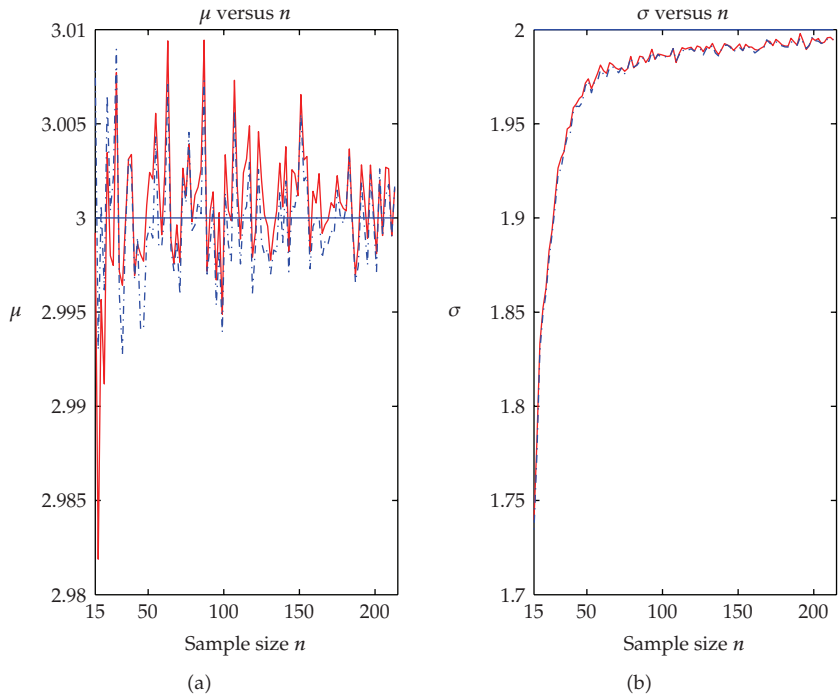

Figure 5: Grouped data: $\tau_1 = -2$, $\tau_2 = 0$, $\tau_3 = 2$, $\tau_4 = 4$, $\tau_5 = 6$. Type I left censored data: $\tau = -2$.

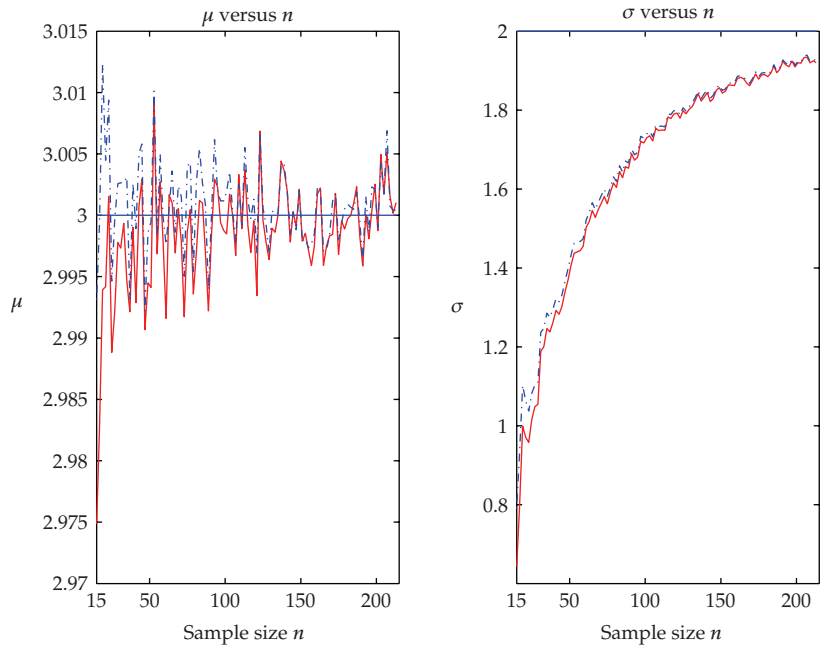

Figure 6: Grouped data: $\tau_1 = -3.5$, $\tau_2 = -0.5$, $\tau_3 = 2.5$, $\tau_4 = 5.5$, $\tau_5 = 8.5$. Type I left censored data: $\tau = -3.5$.

Table 1: Grouped data: $\tau_1 = 2$, $\tau_2 = 2.5$, $\tau_3 = 3$, $\tau_4 = 3.5$, $\tau_5 = 4$.

S.S		$\widehat{m}$			$\widehat{v}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.992	33.374	90.9%	245.490	1517.674	70.7%
35	21.665	31.882	91.7%	229.715	2810.490	78.8%
40	21.441	29.207	92.9%	231.326	2356.412	81.4%
50	21.211	25.697	93.4%	222.684	1632.301	82.9%
100	20.614	17.983	94.5%	199.014	647.065	85.6%

Table 2: Grouped data: $\tau_1 = 1.5$, $\tau_2 = 2.5$, $\tau_3 = 3.5$, $\tau_4 = 4.5$, $\tau_5 = 5.5$.

S.S		$\widehat{m}$			$\widehat{v}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.660	34.091	92.4%	245.569	1835.284	85.4%
35	21.320	30.027	92.1%	211.600	794.547	81.9%
40	21.159	27.386	93.4%	205.518	763.310	84.1%
50	20.884	24.588	94.4%	190.732	525.019	83.3%
100	20.465	16.790	94.3%	167.344	343.435	90.0%

and σ under six different partition sets $\{\tau_1 < \tau_2 < \tau_3 < \tau_4 < \tau_5\}$ with n ranges from 15 to 215. The results of median and mean of the lognormal population are listed in Tables 1, 2, 3, 4, 5, and 6.

Table 3: Grouped data: $\tau_1 = 0$, $\tau_2 = 1.5$, $\tau_3 = 3$, $\tau_4 = 4.5$, $\tau_5 = 6$.

S.S		$\widehat{m}$			$\widehat{v}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.657	31.826	92.7%	203.855	845.742	82.2%
35	21.397	29.168	93.4%	189.163	679.646	83.8%
40	21.201	26.989	93.1%	181.097	558.176	84.6%
50	20.974	23.931	93.6%	172.924	429.353	85.1%
100	20.495	16.577	94.3%	159.906	274.166	89.2%

Table 4: Grouped data: $\tau_1 = -1.5$, $\tau_2 = 1$, $\tau_3 = 3.5$, $\tau_4 = 6$, $\tau_5 = 8.5$.

S.S		$\widehat{m}$			$\widehat{v}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.752	32.374	92.4%	176.952	610.020	80.1%
35	21.381	29.635	92.5%	169.389	512.877	81.2%
40	21.247	28.552	92.6%	168.496	552.496	85.0%
50	21.179	24.738	93.8%	166.919	373.012	83.3%
100	20.507	17.132	94.3%	156.027	264.165	88.8%

Table 5: Grouped data: $\tau_1 = -2$, $\tau_2 = 0$, $\tau_3 = 2$, $\tau_4 = 4$, $\tau_5 = 6$.

S.S		$\widehat{m}$			$\widehat{v}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.688	32.743	93.1%	183.834	659.789	79.8%
35	21.496	29.564	92.3%	180.633	569.356	82.7%
40	21.141	27.514	92.5%	175.077	554.142	84.1%
50	21.093	24.082	93.1%	170.522	360.508	82.7%
100	20.562	16.932	94.3%	158.738	266.687	89.1%

From these figures (grouped data: solid line, type I censoring: dotted line), it is easy to see that estimations under both data situations improved dramatically with the increasing sample size. The estimated values are very close to the true values with error less than 0.003% when $n > 30$. The choice of τ's does not seem to affect the result much except in Figure 6, where $\tau_1 = -3.5$, $\tau_2 = -0.5$, $\tau_3 = 2.5$, $\tau_4 = 5.5$, $\tau_5 = 8.5$, an interval which most samples will be observed in the middle and few on the either side. From those figures, it is not hard to see that the estimation with grouped data are uniformly better than those based on type I censoring data, especially in the estimation of σ, with exception in few isolated cases. Moreover, it is interesting to observe how the $\widehat{\mu}$ and $\widehat{\sigma}$ approach the true value differently with μ taking the oscillated routine and σ tends to be consistently underestimated.

4. An Application

Let us consider a sample of 47 observations from the guidance document USEPA [10, pages 6.22–6.25]. The data describe the measures of 1,2,3,4-Tetrachlorobenzene (TcCB) concentrations (in parts per billion, usually abbreviated ppb) from soil samples at a "Reference" site.

Table 6: Grouped data: $\tau_1 = -3.5$, $\tau_2 = -0.5$, $\tau_3 = 2.5$, $\tau_4 = 5.5$, $\tau_5 = 8.5$.

S.S		$\widehat{m}$			$\widehat{\nu}$	
n	Average	A.W.	C.R.	Average	A.W.	C.R.
30	21.641	32.122	90.8%	170.705	505.267	77.8%
35	21.386	30.498	92.8%	167.457	517.091	80.3%
40	21.250	28.960	93.5%	164.214	555.646	83.9%
50	21.035	26.623	95.2%	161.572	583.249	88.7%
100	20.469	17.481	94.1%	154.492	271.900	88.7%

Table 7: Grouped data: $\tau_1 = -0.71$, $\tau_2 = -0.61$, $\tau_3 = -0.41$, $\tau_4 = -0.21$, $\tau_5 = 0.11$.

n	95% CI for μ	95% CI for σ	95% CI for m	95% CI for ν
47	$(-0.771, -0.426)$	$(0.354, 0.709)$	$(0.455, 0.645)$	$(0.525, 0.741)$

The normal Q-Q plot for the log-transformed TcCB data shown in the book of Millard and Neerchal (2001) indicates that the lognormal distribution appears to provide a good fit to the original data. The book gives $\widehat{\nu}(c) = 0.60$ as the MLE of the mean of the lognormal distribution, and $\mathrm{CI}(c) = [0.51, 0.68]$ as an approximate 95% confidence interval for ν based on the complete sample data with the 47 observations. The book also uses 0.5 as the detection limit, that is, any observation lower than 0.5 will be censored, which yields 19 censored observations and 28 uncensored observations. The censored data then give $\widehat{\nu}(I) = 0.606$ as the MLE of ν and $\mathrm{CI}(I) = [0.51, 0.73]$ as an approximate 95% confidence interval for ν.

To apply the results in Section 2 for computing the MLEs of the parameters of this lognormal distribution, we first transform the original data to their logarithms and thus the log-transformed data constitute a sample from a normal distribution, then obtain $n_1 = 19$, $n_2 = 5$, $n_3 = 7$, $n_4 = 6$, $n_5 = 5$, $n_6 = 5$ by using the following five partition points $\tau_1 = -0.71$, $\tau_2 = -0.61$, $\tau_3 = -0.41$, $\tau_4 = -0.21$, $\tau_5 = 0.11$. Solving the corresponding log-likelihood equations gives $\widehat{\mu}(g) = -0.599$, $\widehat{\sigma}(g) = 0.532$, $\widehat{m}(g) = 0.549$, and $\widehat{\nu}(g) = 0.633$. Approximate 95% confidence intervals for μ, σ, m, and ν are given in Table 7.

Appendix

Proof of Lemma 2.1. To prove the lemma, it is sufficient to verify the following three limits:

$$\lim_{\theta_2 \to 0_+} \sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) = -\infty; \tag{A.1}$$

$$\lim_{\theta_2 \to \infty} \sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) = -\infty; \tag{A.2}$$

$$\lim_{|\theta_1| \to \infty} \sup_{\theta_2>0} \ln L(\theta_1, \theta_2) = -\infty. \tag{A.3}$$

To see (A.1), from the assumption $n_1 + n_{k+1} < n$, there exists an index, say i, such that $2 \le i \le k$ and $n_i > 0$. We have $\ln L(\theta_1, \theta_2) \le n_i \ln \int_{\theta_2\tau_{i-1}-\theta_1}^{\theta_2\tau_i-\theta_1} \varphi(t)dt \le n_i \ln[\theta_2(\tau_i - \tau_{i-1})\varphi(0)]$. So $\sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) \le n_i[\ln\varphi(0) + \ln(\tau_i - \tau_{i-1}) + \ln\theta_2]$ and $\limsup_{\theta_2 \to 0_+} \sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) \le \lim_{\theta_2 \to 0_+} n_i[\ln\varphi(0) + \ln(\tau_i - \tau_{i-1}) + \ln\theta_2] = -\infty$. Therefore, (A.1) holds.

To show (A.2), we denote $I \equiv \{1 \le j \le k+1,\ n_j > 0\}$. For each fixed $\theta_2 > 0$, it is evident that $\ln L(\theta_1, \theta_2) = \sum_{i\in I} n_i \ln \int_{\theta_2\tau_{i-1}-\theta_1}^{\theta_2\tau_i-\theta_1} \varphi(t)dt \equiv M(\theta_2)$. Thus $\sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) = \sup_{-\infty<\theta_1<\infty} M(\theta_2)$.

Note that, $\lim_{|\theta_1|\to\infty} M(\theta_1) = -\infty$, so there exists $\theta_1^* = \theta_1^*(\theta_2) \in (-\infty, \infty)$, such that $\sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) = \ln L(\theta_1^*, \theta_2) = \sum_{i\in I} n_i \ln \int_{\theta_2\tau_{i-1}-\theta_1^*}^{\theta_2\tau_i-\theta_1^*} \varphi(t)dt$. Consider function $g(x) \equiv |x| \exp(x^2/2)$. For any given large number $A > 0$, it is easy to see that there exists $x_0 > 0$ such that $g(x) > \sqrt{2\pi} \exp(x^2/2),\ \forall |x| > x_0$.

Denote $c \equiv \min_{1\le j\le k+1}(\tau_j - \tau_{j-1}) > 0$. For any $\theta_2 > x_0/c$, from our assumptions there exists an index, say i, belonging to I satisfying (a) $n_i > 0$; (b) the following two quantities $\theta_2\tau_{i-1} - \theta_1^* = \theta_2(\tau_{i-1} - \theta_1^*/\theta_2)$ and $\theta_2\tau_i - \theta_1^* = \theta_2(\tau_i - \theta_1^*/\theta_2)$ have the same sign; and (c) $|\tau_{i-1} - \theta_1^*/\theta_2| > c$, and $|\tau_i - \theta_1^*/\theta_2| > c$.

Note that, if $i \in I$, and both $\theta_2\tau_{i-1} - \theta_1^* > 0$ and $\theta_2\tau_i - \theta_1^* > 0$, then

$$\begin{aligned}
\int_{\theta_2\tau_{i-1}-\theta_1^*}^{\theta_2\tau_i-\theta_1^*} \varphi(t)dt &< \frac{1}{\theta_2\tau_{i-1} - \theta_1^*} \int_{\theta_2\tau_{i-1}-\theta_1^*}^{\theta_2\tau_i-\theta_1^*} t\varphi(t)dt \\
&= \frac{1}{\theta_2\tau_{i-1} - \theta_1^*} \left[\varphi\left(\theta_2\tau_{i-1} - \theta_1^*\right) - \varphi\left(\theta_2\tau_i - \theta_1^*\right)\right] \\
&< \frac{\varphi\left(\theta_2\tau_{i-1} - \theta_1^*\right)}{\theta_2\tau_{i-1} - \theta_1^*} = \frac{1}{\sqrt{2\pi}\ \ g\left(\theta_2\tau_{i-1} - \theta_1^*\right)}.
\end{aligned} \tag{A.4}$$

If $\theta_2 > x_0/c$, then $\theta_2\tau_{i-1} - \theta_1^* = \theta_2(\tau_{i-1} - \theta_1^*/\theta_2) > (x_0/c)c = x_0$ and so $g(\theta_2\tau_{i-1} - \theta_1^*) > \sqrt{2\pi} \exp A$. Consequently,

$$n_i \ln \int_{\theta_2\tau_{i-1}-\theta_1^*}^{\theta_2\tau_i-\theta_1^*} \varphi(t)dt < n_i \ln \frac{1}{\sqrt{2\pi} \exp(A)} = n_i\left(-\ln\sqrt{2\pi} - A\right) < -n_i A < -A. \tag{A.5}$$

This further implies

$$\begin{aligned}
\sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) = \ln L\left(\theta_1^*, \theta_2\right) &= \sum_{j\in I} n_j \ln \int_{\theta_2\tau_{j-1}-\theta_1^*}^{\theta_2\tau_j-\theta_1^*} \varphi(t)dt \\
&< n_i \ln \int_{\theta_2\tau_{i-1}-\theta_1^*}^{\theta_2\tau_i-\theta_1^*} \varphi(t)dt < -A, \quad \forall \theta_2 > \frac{x_0}{c}.
\end{aligned} \tag{A.6}$$

If $i \in I$, but both $\tau_{i-1} - \theta_1^*/\theta_2 < -c$ and $\tau_i - \theta_1^*/\theta_2 < -c$, then similarly, it can be shown that (A.6) is true again. Therefore, we see that for any given large number $A > 0$, it holds that

$$\sup_{-\infty<\theta_1<\infty} \ln L(\theta_1, \theta_2) < -A, \quad \forall \theta_2 > \frac{x_0}{c}. \tag{A.7}$$

Due to the arbitrariness of $A > 0$, we conclude that (A.1) is true.

To verify (A.3), we let $\theta_2\tau_{k+1} - \theta_1 = \infty$ and $\theta_2\tau_0 - \theta_1 = -\infty$ for any $(\theta_1, \theta_2) \in (-\infty, \infty) \times (0, \infty)$. For any fixed $\theta_1 \in (-\infty, \infty)$, we have

$$\ln L(\theta_1, \theta_2) = \sum_{j \in I} n_j \ln \int_{\theta_2\tau_{j-1}-\theta_1}^{\theta_2\tau_j-\theta_1} \varphi(t)dt \equiv M(\theta_2). \tag{A.8}$$

It can be easily verified that $M(\theta_2) \to -\infty$ as $\theta_2 \to 0+$ or $\theta_2 \to \infty$. Thus, there exists $\theta_2^* \equiv \theta_2^*(\theta_1) \in (0, \infty)$ such that $\sup_{\theta_2>0} \ln L(\theta_1, \theta_2) = \ln L(\theta_1, \theta_2^*)$.

We define function $g(x)$, x_0 for any given $A > 0$, and $c > 0$ as before. Consider any sequence $\{\theta_{1m},\ m \geq 1\} \subset (-\infty, \infty)$ with $|\theta_{1m}| \to \infty$ as $m \to \infty$. Let $\theta_{2m}^* \equiv \theta_2^*(\theta_{1m})$ and $\{\theta_{2m_r}^*,\ r \geq 1\}$ be any converging subsequence of $\{\theta_{2m},\ m \geq 1\}$, $\eta \equiv \lim_{m\to\infty}\theta_{2m_r}^* \leq \infty$. Let us study two cases.

Case 1 ($\eta = \infty$). Notice that for any $r \geq 1$, by our assumptions there exists at least one index, say i, in I such that (a) $n_i > 0$; (b) $|\tau_{i-1} - \theta_{1m_r}/\theta_{2m_r}^*| > c$ and $|\tau_i - \theta_{1m_r}/\theta_{2m_r}^*| > c$; (c) $\tau_{i-1} - \theta_{1m_r}/\theta_{2m_r}^*$ and $\tau_i - \theta_{1m_r}/\theta_{2m_r}^*$ have the same sign.

Since $\theta_{2m}^* \to \infty$ as $m \to \infty$, there exists r_0 sufficiently large such that $\theta_{2m_r}^* > x_0/c,\ \forall r \geq r_0$. Thus,

$$\begin{aligned} \left|\theta_{2m_r}^*\tau_{i-1} - \theta_{1m_r}\right| &= \theta_{2m_r}^*\left|\tau_{i-1} - \frac{\theta_{1m_r}}{\theta_{2m_r}^*}\right| > x_0, \quad \forall r \geq r_0, \\ \left|\theta_{2m_r}^*\tau_i - \theta_{1m_r}\right| &= \theta_{2m_r}^*\left|\tau_i - \frac{\theta_{1m_r}}{\theta_{2m_r}^*}\right| > x_0, \quad \forall r \geq r_0. \end{aligned} \tag{A.9}$$

From these, as what we did before we obtain

$$\ln L\left(\theta_{1m_r}, \theta_{2m_r}^*\right) < n_i \ln \int_{\theta_{2m_r}^*\tau_{i-1}-\theta_{1m_r}}^{\theta_{2m_r}^*\tau_i-\theta_{1m_r}} \varphi(t)dt < -A, \quad \forall r \geq r_0, \tag{A.10}$$

this implies $\lim_{r\to\infty} \ln L(\theta_{1m_r}, \theta_{2m_r}^*) = -\infty$.

Case 2 ($0 \leq \eta < \infty$). In this case, the inequality $\lim_{r\to\infty} \ln L(\theta_{1m_r}, \theta_{2m_r}^*) = -\infty$ can be proved in the same way as Case 1.

From the results in the above two cases, we conclude that $\lim_{m\to\infty} \ln L(\theta_{1m}, \theta_{2m}^*) = -\infty$. Since $\{\theta_{1m},\ m \geq 1\}$ is an arbitrary sequence satisfying $|\theta_{1m}| \to \infty$, so finally (A.3) is true. □

Proof of Lemma 2.2. For any given $v < u$, we have $g(u, v) \equiv \ln(\Phi(u) - \Phi(v))$. The Hessian matrix of $g(u, v)$ is

$$H^* = \begin{pmatrix} \dfrac{\varphi'(u)(\Phi(u) - \Phi(v)) - \varphi^2(u)}{(\Phi(u) - \Phi(v))^2} & \dfrac{\varphi(u)\varphi(v)}{(\Phi(u) - \Phi(v))^2} \\ \dfrac{\varphi(u)\varphi(v)}{(\Phi(u) - \Phi(v))^2} & -\dfrac{\varphi'(v)(\Phi(u) - \Phi(v)) + \varphi^2(v)}{(\Phi(u) - \Phi(v))^2} \end{pmatrix}. \tag{A.11}$$

In order to prove H^* is negative definite, the following two conditions must be satisfied: (i) $\partial^2 g/\partial u^2 < 0$ or $\partial^2 g/\partial v^2 < 0$; (ii) the determinant of the Hessian matrix H^* is positive.

The inequality $\partial^2 g/\partial v^2 < 0$ is equivalent to $\varphi(v) > v(\Phi(u) - \Phi(v))$. This inequality follows from $y[1 - \Phi(y)] < \phi(y),\ \forall y$. Thus the desired inequality is true.

From the expression of H^*, it follows that

$$\begin{aligned}&(\Phi(u) - \Phi(v))^2 |H^*| \\ &\quad = -\varphi'(u)\varphi'(v)(\Phi(u) - \Phi(v))^2 - \varphi'(u)\varphi^2(v)(\Phi(u) - \Phi(v)) + \varphi^2(u)\varphi'(v)(\Phi(u) - \Phi(v)).\end{aligned} \tag{A.12}$$

The inequality $|H^*| > 0$ is equivalent to $u\varphi(v) - v\varphi(u) - uv(\Phi(u) - \Phi(v)) > 0$. We discuss three cases.

Case 1 ($v < u \le 0$). We have $-u(\Phi(u) - \Phi(v)) = \int_v^u -u\varphi(t)dt < \int_v^u -t\varphi(t)dt = \varphi(u) - \varphi(v)$. From this, we see that

$$\begin{aligned}&u\varphi(v) - v\varphi(u) - uv(\Phi(u) - \Phi(v)) \\ &\quad > u\varphi(v) - v\varphi(u) + v\big[\varphi(u) - \varphi(v)\big] = (u - v)\varphi(v) > 0.\end{aligned} \tag{A.13}$$

Case 2 ($v < 0 < u$). It is obvious that $u\varphi(v) - v\varphi(u) - uv(\Phi(u) - \Phi(v)) > 0$.

Case 3 ($0 < v < u$). It holds that

$$v(\Phi(u) - \Phi(v)) = \int_v^u v\varphi(t)dt < \int_v^u t\varphi(t)dt = \int_v^u -\varphi'(t)dt = \varphi(v) - \varphi(u). \tag{A.14}$$

From this, we see that $-uv(\Phi(u) - \Phi(v)) > -u(\varphi(v) - \varphi(u)) = -u\varphi(v) + u\varphi(u)$ since $u > 0$. It means that $u\varphi(v) - u\varphi(u) - uv(\Phi(u) - \Phi(v)) > 0$. This further implies that $u\varphi(v) - v\varphi(u) - uv(\Phi(u) - \Phi(v)) > 0$ since $u > v > 0$. Hence, in all the three cases, we obtain $|H^*| > 0$.

From all the above, we conclude that both conditions (i) and (ii) are satisfied and thus the Hessian matrix $H^*(u, v)$ is negative definite. □

Notations

S.S.:	Sample size
$\widehat{\mu}(g)$, $\widehat{\sigma}(g)$:	MLEs of μ, σ with grouped data
$\widehat{\mu}(I)$, $\widehat{\sigma}(I)$:	MLEs of μ, σ with type I left censored data
$\widehat{m}$:	MLE of median $m = \exp\{\mu\}$ of $LN(\mu, \sigma^2)$ distribution with grouped data
$\widehat{v}$:	MLE of mean $v = \exp\{\mu + \sigma^2/2\}$ of $LN(\mu, \sigma^2)$ distribution with grouped data
Average:	The average of estimates from 5000 simulations
A.W.:	The average width of 5000 approximate 95% confidence intervals
C.R.:	The average coverage rate of 5000 approximate 95% confidence intervals.

References

[1] P. Tai, E. Yu, R. Shiels, and J. Tonita, "Long-term survival rates of laryngeal cancer patients treated by radiation and surgery, radiation alone, and surgery alone: studied by lognormal and Kaplan-Meier survival methods," *BMC Cancer*, vol. 5, article 13, 2005.

[2] R. F. Mould, M. Lederman, P. Tai, and J. K. M. Wong, "Methodology to predict long-term cancer survival from short-term data using Tobacco Cancer Risk and Absolute Cancer Cure models," *Physics in Medicine and Biology*, vol. 47, no. 22, pp. 3893–3924, 2002.

[3] M. V. Ratnaparkhi and W. J. Park, "Lognormal distribution: model for fatigue life and residual strength of composite materials," *IEEE Transactions on Reliability*, vol. 35, pp. 312–315, 1986.

[4] D. L. Kelly, "Use of constrained lognormal distribution in reliability analysis," *Reliability Engineering & System Safety*, vol. 40, no. 1, pp. 43–47, 1993.

[5] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ, USA, 2003.

[6] J. E. Mosimann and G. Campbell, "Applications in biology of the lognormal distribution: simple growth models," in *Lognormal Distribution: Theory and Applications*, pp. 287–302, 1988.

[7] T. Mäkeläinen, K. Schmidt, and G. P. H. Styan, "On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples," *The Annals of Statistics*, vol. 9, no. 4, pp. 758–767, 1981.

[8] W. Feller, *An Introduction to Probability Theory and Its Applications: Volume 2*, John Wiley & Sons, New York, NY, USA, 1957.

[9] N. Balakrishnan and J. Mi, "Existence and uniqueness of the MLEs for normal distribution based on general progressively type-II censored samples," *Statistics & Probability Letters*, vol. 64, no. 4, pp. 407–414, 2003.

[10] USEPA, *Statistical Methods for Evaluating the Attainment of Cleanup Standards. Vol. 3*, 1994.