RESEARCH ARTICLE

Exponentiated Log-normal (ELN) Processes with Left Censored Data

Cheng Peng,

Department of Mathematics, West Chest University of Pennsylvania, West Chester, PA 19383, USA.

ARTICLE HISTORY
Compiled October 11, 2023

This work is in progress. Please do not cite or quote without the author's permission.

Contents

Address of correspondence: C. Peng. Email: cpeng@wcupa.edu

ABSTRACT
In this paper, we propose to fit exponentiated lognormal (ELN) distribution to environmental data with multiple detection limits. Since exponentiated lognormal distribution is a generalization of the usual lognormal distribution, some stochastic and numerical comparisons between the two distributions will be summarized. The standard maximum likelihood estimation is used to estimate the population parameters. The large sample inference mainly focuses on population mean and testing on necessity of the resilience parameter. Guidelines of selecting either asymptotic or bootstrap confidence intervals for practitioners are provided. We also provide some illustrative examples using real-life data.

KEYWORDS
Exponentiated lognormal distribution, asymptotic confidence interval, quantile estimation, Pearson χ^2 goodness-of-fit test, detection limit.

1. Introduction

Data with non-detect values are common in many areas such as environmental sciences in which instruments cannot detect measurements below certain levels. Various estimation methods have been proposed to handle such data in analysis. [16] and [22] provide a good review on commonly used existing methods. The recent monograph by [21] gives a detailed discussion on various methods available for analyzing data with detection limits.

One step replacement methods have been proposed and used for decades. These methods are usually produce poor results or even ruin the results, see [23]. The likelihood based methods are the major statistical tools in analyzing these data. The maximum likelihood estimation (MLE) has been used in analyzing this day since 1950s by researcher including [18], [1], [18], [2], and among others. The monograph by [3] systematically introduces various methods analyzing left censored and truncated data. [30] developed an asymptotic interval and bootstrap intervals for population means and variance based on left censored data with multiple limits.. Uses the bootstrap sampling distribution to decide whether to trust the asymptotic interval.

The *Expectation and Maximization* (EM) algorithm proposed by [5] is an iterated algorithm for finding the MLE of the model parameters. The idea of EM algorithm is to use the completely observed data to estimate model parameters and then use these estimated parameters to predict the censored values, and then repeat the same procedure until convergence of the estimation is achieved. In other words, the EM algorithm can be viewed as a model based iterative replacement method for find the MLE. This interesting observation was also described in [13].

Some nonparametric approaches have also been discussed in literature. Reader are referred to, for example, [20], [16], [12], and [33].

Since environmental data are positive and usually skewed, the parametric distributions that have been used for non-detect data are log-normal, gamma and Weibull, see [36]. The most popular model used in practice is the log-normal distribution. The normal distribution has been widely use in practice since lognormal data can be transformed to normal data. In this paper we propose to fit exponentiated log-normal distribution to skew environmental data. Exponentiated lognormal distribution includes lognormal as a special case. Its cumulative probability distribution function (CDF) is the power of the CDF of lognormal. That is,

$$G(x) = \left[\Phi\left(\frac{\log(x) - \xi}{\eta}\right)\right]^{\alpha}, \tag{1}$$

where α is the new parameter added to the CDF of lognormal distribution with location parameter ξ and scale parameter η. $\Phi(\cdot)$ is the CDF of the standard normal distribution. In fact, it is a member of more general family of distributions discussed in [24]. α is called the resilience parameter which is also called proportional reverse hazard rate in lifetime survival analysis. [19] call this model log-power-normal distribution and use it to fit air pollution data. We will use this model to fit environmental data with detection limits.

The rest of the paper is organized as follows. In Section 2, we briefly present some results on comparison between the lognormal and exponentiated lognormal distribution. The maximum likelihood estimation of the model parameters, expectation of the distribution and asymptotic confidence intervals in Sections 3 and 4. The The confidence interval of quanitle estimate is given in Section 5. In Section 6, we discuss the large sample Pearson χ^2 test for the goodness-of-fit of the model. Several illustrative examples using real-life environmental data sets and some guidelines for selecting appropriate confidence intervals are presented in Section 7. Discussions and concluding remarks are given in Section 8.

2. Properties and Visual Analyses of ELN

We first introduce some notations and definitions of stochastic orders. Denote the CDF and PDF of lognormal distribution to be

$$F(x) = \Phi\left(\frac{\log(x) - \xi}{\eta}\right) \quad \text{and} \quad f(x) = \frac{1}{x\eta}\phi\left(\frac{\log(x) - \xi}{\eta}\right) \tag{2}$$

where ξ and η are location and scale parameters respectively. We also denote the density function of the exponentiated lognormal distribution to be

$$g(x) = \frac{\alpha}{x\eta}\phi\left(\frac{\log(x) - \xi}{\eta}\right)\left[\Phi\left(\frac{\log(x) - \xi}{\eta}\right)\right]^{\alpha - 1} \tag{3}$$

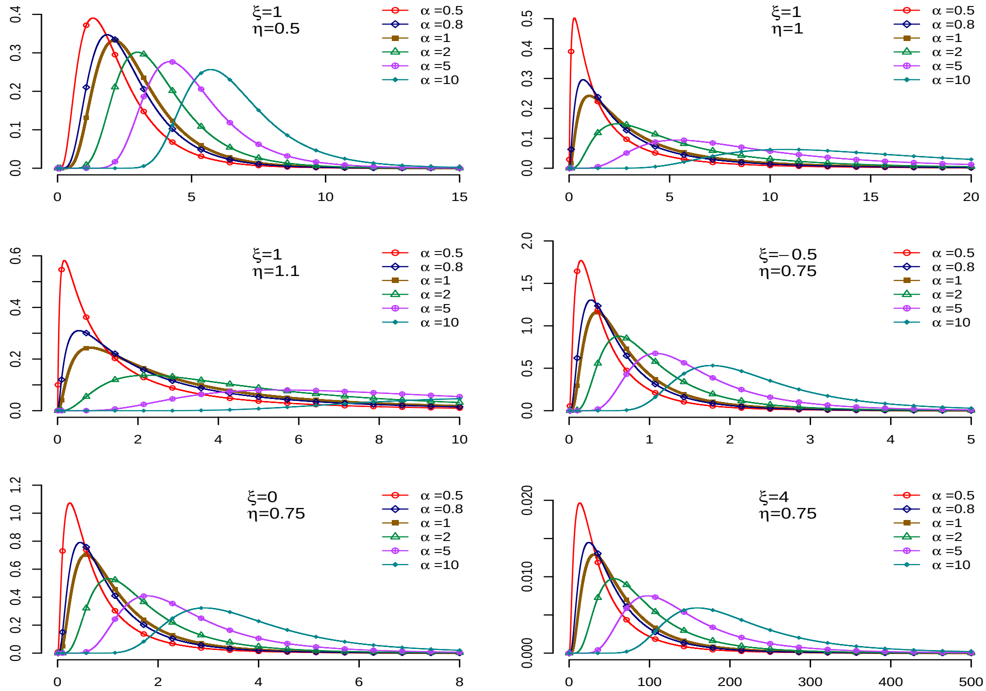

Figure 1. Density curves of exponentiated lognormal distributions with various values of resilience parameter α, the location and scale parameters of the reference lognormal distribution ξ and η, respectively.

We observe several patterns from Figure 1. First of all, all density curves skewed to the right (positively skewed) regardless the choice of values of the three parameters. In addition,

1. As the value of resilience parameter increases, the corresponding curve becomes flatter for fixed ξ and η. This implies that both expected and variance increase as the resilience parameter increases.

2. If $\alpha < 1$, both of the mean and variance of exponentiated lognormal distribution are less than the lognormal distribution. However, both of the mean and variance of exponentiated lognormal distribution are greater than the lognormal distribution.

Let $X \rightharpoonup F(x) = \Phi[\log(x) - \xi]/\eta)$ and $Y \rightharpoonup G(x)$. Following result gives the relationship between X and Y.

Theorem 2.1. *Let Y and X be two random variables having distributions in (1) and (2). We have* $P(Y > X) = \alpha/(1+\alpha)$.

Proof. We use ROC curve method to show this result. Let $\bar{G}(x)$ and $\bar{F}(x)$ be the survival function of Y and X. From chapter 4 of [27], the ROC curve is defined to be $[\bar{F}(x), \bar{G}(x)]$ for all $x \in \mathbb{R}^{+1}$. Let $t = \bar{F}(x) = 1 - \Phi[(\log(x) - \xi)\eta]$. Then $\Phi^{-1}(1-t) = (\log(x) - \xi)/\eta$. The ROC curve is then defined as

$$\text{ROC}(t) = 1 - \left[\Phi\left(\frac{\log(x) - \xi}{\eta}\right)\right]^{\alpha} = 1 - \left[\Phi\left(\Phi^{-1}(1-t)\right)\right]^{\alpha} = 1 - (1-t)^{\alpha}. \tag{4}$$

From Result 4.2 of [27] [page 69], we have

$$P(Y > X) = \int_0^1 \text{ROC}(t)dt = \int_0^1 [1 - (1-t)^{\alpha}]dt = \frac{\alpha}{1+\alpha}. \tag{5}$$

The proof is complete. □ □

The key numerical characteristics of interest in this paper are the expectation and the variance of the exponentiated distribution. The next result concerns about the monotonicity of r-th moment of Y. Note that the r-th moment of Y is defined to be

$$E[Y^r] = \int_0^\infty y^r d\left[\Phi\left(\frac{\log(y)-\xi}{\eta}\right)\right]^\alpha. \tag{6}$$

After changing variable

$$x = \left[\Phi\left(\frac{\log(y)-\xi}{\eta}\right)\right]^\alpha, \quad \text{that is,} \quad y = \exp\left[\xi + \eta\Phi^{-1}(x^{\frac{1}{\alpha}})\right].$$

We re-express the r-th moment as

$$E[Y^r] = \int_0^1 \exp\left\{r\left[\xi + \eta\Phi^{-1}(x^{\frac{1}{\alpha}})\right]\right\} dx. \tag{7}$$

Let $t = x^{1/\alpha}$, then $dx = \alpha t^{\alpha-1}$. With this change of variable, we re-express the t-th moment in the following

$$\mu_Y^{(r)} = \int_0^1 \alpha t^{\alpha-1} \exp[r\xi + r\eta\Phi^{-1}(t)]dt \tag{8}$$

Therefore, the mean and variance of the exponentiated lognormal distribution are given by

$$\mu = E[Y] = \int_0^1 \alpha t^{\alpha-1} \exp\left[\xi + \eta\Phi^{-1}(t)\right] dt \tag{9}$$

and

$$\sigma^2 = \mathrm{V}(Y) = \int_0^1 \alpha t^{\alpha-1} \exp\left\{2\left[\xi + \eta\Phi^{-1}(t)\right]\right\} dt - \mu^2 \tag{10}$$

To highlight the functional relationship between the moment and the resilience parameter α, we denote the write $\mu_Y^{(r)}(\alpha) = E[Y^r]$.

Theorem 2.2. *Under model (1), the r-th moment of the exponentiated lognormal distribution is an increasing function in α for fixed ξ and η.*

Proof. We only show the first order derivative is positive for all $\alpha \in \mathbb{R}^+$. Observe that

$$\frac{d}{d\alpha}\mu_Y^{(r)}(\alpha) = \int_0^1 \exp\left\{r\left[\xi + \eta\Phi^{-1}(x^{\frac{1}{\alpha}})\right]\right\} \cdot r\eta\frac{d\Phi^{-1}(x^{1/\alpha})}{d\alpha}dx. \tag{11}$$

Denote $\psi = \Phi^{-1}(x^{1/\alpha})$. We solve for x and obtain $[\Phi(\psi)]^\alpha = x$, which is equivalent to $\alpha\log[\Phi(\psi)] = \log(x)$. Taking derivative of implicit function, we have

$$\frac{d\psi}{d\alpha} = \frac{-x^{1/\alpha}\log x^{1/\alpha}}{\alpha\phi[\Phi^{-1}(x^{1/\alpha})]} > 0 \tag{12}$$

since $0 < x < 1$, $\log(x) < 0$. Substituting $d\psi/d\alpha$ in (12) to (11) and notice that the integrand in (11) is positive. We conclude that $\frac{d}{d\alpha}\mu_Y^r(\alpha) > 0$. Therefore, the r-th moment is increasing in α. □

Corollary 2.3. *For fixed location and scale parameters in the exponentiated lognormal distribution, the expectation is an increasing functions of the resilience parameter α.*

Proof. The monotonicity of the expectation has been proved in Theorem 2. Next we show that the variance of the exponentiated lognormal is increasing in α. In the proof of Theorem 2, we have have shown that $\psi = \Phi^{-1}(x^{1/\alpha})$ in increasing in α. Similar arguments will yield that ψ is also increasing in x for $x \in [0,1]$. □

In stead of presenting the cumbersome derivations, we use the following figure to visualize how the values of parameters impact the shape the variance

The above Figure 2 shows that, when fixing any two of the three parameters, the variance as a function of the third parameter is an increasing function. This is consistent with the part of the

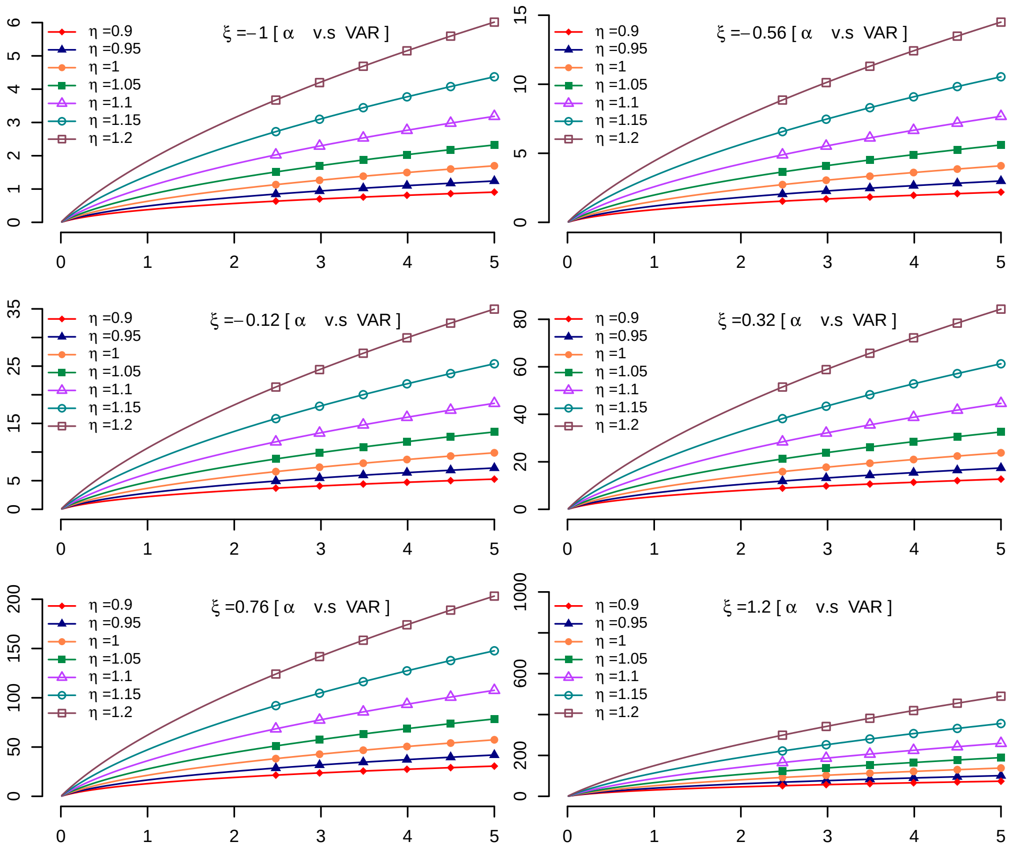

Figure 2. Variances of exponentiated lognormal distributions with various values of resilience parameter α, the location and scale parameters of the reference lognormal distribution ξ and η, respectively.

information observed in Figure 1. We also see from Figure 1 that exponentiated lognormal distributions are positively skewed. The following Figure 3 confirms the postive skewness with some additional information about how individual parameters impact the degrees of skewness.

First of all, Figure 3 shows that the coefficient of skewness is always positive (the values of parameters were chosen to demonstrate the overall view and avoid the overflow error and divergence issue in the operations). We can see from the top three panels that, for fixed ξ and η, the skewness coefficient is a decreasing function of α. This is consistent with the pattern observed in Figure 1. However, The coefficient of skewness of the exponentiated lognormal is an increasing of η for fixed α and ξ. The degrees of skewness seem not not to impacted significantly by ξ.

3. Maximum Likelihood Estimation and Asymptotic Results

Let $\{y_1, \cdots, y_n\}$ be the type I left censored sample taken from a normal population $N(\xi, \eta^2)$ with $y_i = (x_i, c_i)$ where c_i is the censoring variable taking value 1 if x_i is censored and 0 if x_i is completely observed. From (1) and (3), we can write the likelihood function based on the data with left censored values as

$$\mathcal{L}(\xi,\eta,\alpha) = \prod_{i=1}^{n} \left[\Phi\left(\frac{\log(x_i)-\xi}{\eta}\right)\right]^{\alpha c_i} \left\{\frac{\alpha}{x_i\eta}\phi\left(\frac{\log(x_i)-\xi}{\eta}\right)\left[\Phi\left(\frac{\log(x_i)-\xi}{\eta}\right)\right]^{\alpha-1}\right\}^{1-c_i}, \tag{13}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and density function of the standard normal distribution. ξ, η and α) are the parameters of underlying exponentiated log-normal population that are referred to location, scale and shape parameters respectively. Since type I left censored samples have fixed censoring limits, to simplify the notation, we assume that $\{x_1, x_2, \cdots, x_{n_0}\}$ is the set of complete observations and the fixed number multiple detection limits are given by $(d_1, d_2, \cdots, d_k)$ with corresponding censoring counts $m_1, m_2, \cdots, m_k$.

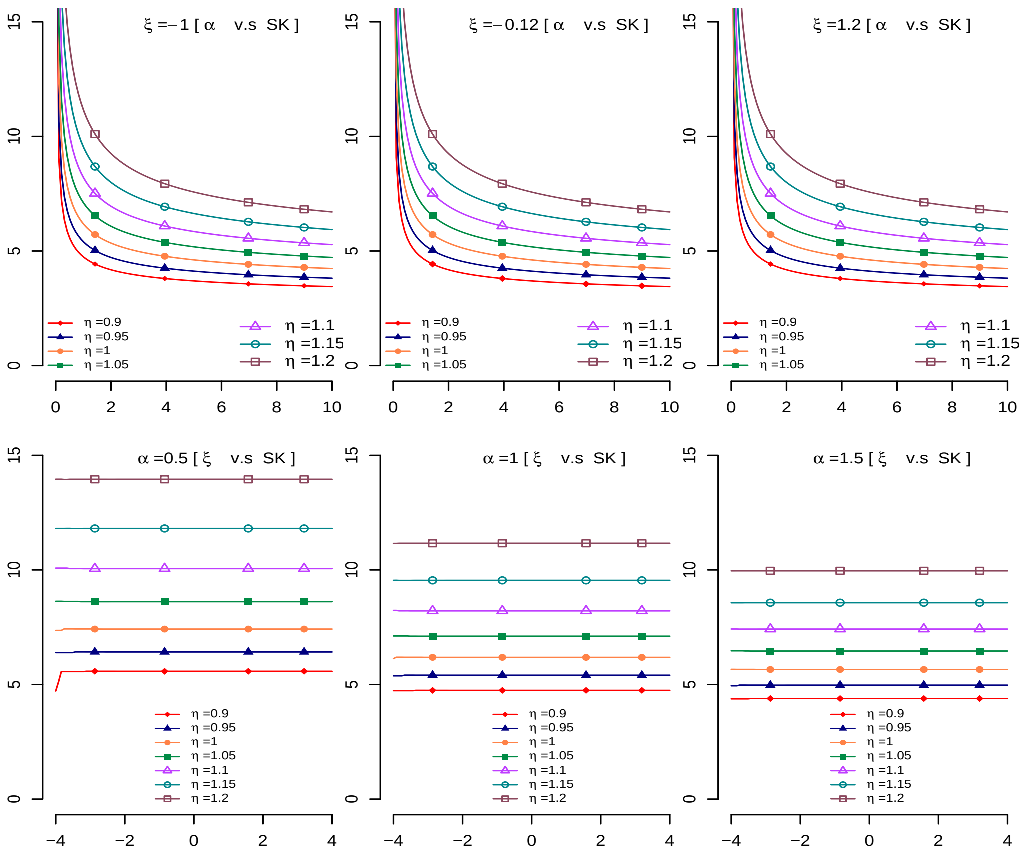

Figure 3. Variances of exponentiated lognormal distributions with various values of resilience parameter α, the location and scale parameters of the reference lognormal distribution ξ and η, respectively.

The corresponding kernel of the log-likelihood based on (13) is given by

$$\updownarrow(\xi,\eta,\alpha)=\alpha\sum_{j=1}^{k}m_j\ln\Phi\left(\frac{\log(d_j)-\xi}{\eta}\right)+(\alpha-1)\sum_{i=1}^{n_0}\ln\Phi\left(\frac{\log(x_i)-\xi}{\eta}\right)$$

$$+n_0(\log\alpha-\ln\eta)-\frac{1}{2}\sum_{i=1}^{n_0}\left(\frac{\log(x_i)-\xi}{\eta}\right)^2. \tag{14}$$

The maximum likelihood estimator of (ξ,η,α), denoted by $(\hat{\xi},\hat{\eta},\hat{\alpha})$, is the solution to the following optimization problem

$$(\hat{\xi},\hat{\eta},\hat{\alpha})=\sup\max_{\xi,\eta,\alpha}\updownarrow(\xi,\eta,\alpha). \tag{15}$$

To derive the score equations and the Hessian matrix for the numerical solution. We need to simplify the notations. Denote, for $j=1,2,\cdots,k$ and $i=1,2,\cdots,n_0$

$$\gamma_j=\frac{\log(d_j)-\xi}{\eta}\text{ and }z_j=\frac{\log(x_j)-\xi}{\eta}. \tag{16}$$

With the above notations, the kernel of the log-likelihood function (14) can be re-expressed in the following

$$\updownarrow(\xi,\eta,\alpha)=\alpha\sum_{j=1}^{k}m_j\ln\Phi\left(\gamma_j\right)+(\alpha-1)\sum_{i=1}^{n_0}\ln\Phi\left(z_i\right)+n_0(\log\alpha-\ln\eta)-\frac{1}{2}\sum_{i=1}^{n_0}z_i^2. \tag{17}$$

Since γ_j and z_i are functions of ξ and η, the partial derivatives of γ_j and z_i are given by

$$\frac{\partial \gamma_j}{\partial \xi} = \frac{\partial z_j}{\partial \xi} = -\frac{1}{\eta}, \quad \frac{\partial \gamma_j}{\partial \eta} = -\frac{\log(d_j) - \xi}{\eta^2} = -\frac{\gamma_j}{\eta}, \quad \text{and} \quad \frac{\partial z_i}{\partial \eta} = -\frac{\log(x_i) - \xi}{\eta^2} = -\frac{z_i}{\eta}. \tag{18}$$

Note also that, for $t = [\log(s) - \xi]/\eta$,

$$\frac{\partial \Phi(t)}{\partial \xi} = -\frac{\phi(t)}{\eta}, \quad \frac{\partial \Phi(t)}{\partial \eta} = -\frac{t\phi(t)}{\eta}, \quad \frac{\partial \phi(t)}{\partial \xi} = \frac{t\phi(t)}{\eta}, \quad \text{and} \quad \frac{\partial \phi(t)}{\partial \eta} = \frac{t^2\phi(t)}{\eta}. \tag{19}$$

Therefore, the system of score equations is given in the following

$$\frac{\partial l(\xi,\eta,\alpha)}{\partial \xi} = -\frac{1}{\eta}\left[\sum_{j=1}^{k} \frac{\alpha m_j \phi(\gamma_j)}{\Phi(\gamma_j)} + \sum_{i=1}^{n_0} \frac{(\alpha-1)\phi(z_i)}{\Phi(z_i)} - \sum_{i=1}^{n_0} z_i\right] = 0, \tag{20}$$

$$\frac{\partial l(\xi,\eta,\alpha)}{\partial \eta} = -\frac{1}{\eta}\left[\sum_{j=1}^{k} \frac{\alpha m_j \gamma_j \phi(\gamma_j)}{\Phi(\gamma_j)} + \sum_{i=1}^{n_0} \frac{(\alpha-1)z_i\phi(z_i)}{\Phi(z_i)} + n_0 - \sum_{i=1}^{n_0} z_i^2\right] = 0, \tag{21}$$

$$\frac{\partial l(\xi,\eta,\alpha)}{\partial \alpha} = \sum_{j=1}^{k} m_j \ln \Phi(\gamma_j) + \frac{n_0}{\alpha} + \sum_{i=1}^{n_0} \ln \Phi\left(z_i\right) = 0. \tag{22}$$

Since the solutions to (20)-(22) have no closed form, numerical methods is used to find the solutions to the score equations. The following Hessian matrix is used the he Newton-Raphson method.

$$\mathrm{H} = \begin{pmatrix} \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \xi^2} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \xi \partial \eta} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \xi \partial \alpha} \\ \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \eta \partial \xi} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \eta^2} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \eta \partial \alpha} \\ \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \alpha \partial \xi} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \alpha \partial \eta} & \frac{\partial^2 l(\xi,\eta,\alpha)}{\partial \alpha^2} \end{pmatrix} \equiv \begin{pmatrix} h_{11}(\xi,\eta,\alpha) & h_{12}(\xi,\eta,\alpha) & h_{13}(\xi,\eta,\alpha) \\ h_{21}(\xi,\eta,\alpha) & h_{22}(\xi,\eta,\alpha) & h_{23}(\xi,\eta,\alpha) \\ h_{31}(\xi,\eta,\alpha) & h_{32}(\xi,\eta,\alpha) & h_{33}(\xi,\eta,\alpha) \end{pmatrix} \tag{23}$$

To derive the individual cells in the Hessian matrix, we will use the following partial derivatives, for $t = [\log(s) - \xi]/\eta$,

$$\frac{\partial}{\partial \xi}\left(\frac{\phi(t)}{\Phi(t)}\right) = \frac{1}{\eta}\frac{\phi(t)}{\Phi(t)}\left(t + \frac{\phi(t)}{\Phi(t)}\right) \quad \text{and} \quad \frac{\partial}{\partial \eta}\left(\frac{\phi(t)}{\Phi(t)}\right) = \frac{t}{\eta}\frac{\phi(t)}{\Phi(t)}\left(t + \frac{\phi(t)}{\Phi(t)}\right) \tag{24}$$

After some algebra, we have

$$\begin{aligned}
-h_{11} &= \frac{1}{\eta^2}\left[\sum_{j=1}^{k} \frac{\alpha m_j \phi(\gamma_j)}{\Phi(\gamma_j)}\left(\gamma_j + \frac{\phi(\gamma_j)}{\Phi(\gamma_j)}\right) + \sum_{i=1}^{n_0} \frac{(\alpha-1)\phi(z_i)}{\Phi(z_i)}\left(z_i + \frac{\phi(z_i)}{\Phi(z_i)}\right) + n_0\right], \\
-h_{12} = -h_{21} &= \frac{1}{\eta^2}\left[\sum_{j=1}^{k} \frac{\alpha m_j \phi(\gamma_j)}{\Phi(\gamma_j)}\left(-1 + \gamma_j^2 + \frac{\gamma_j\phi(\gamma_j)}{\Phi(\gamma_j)}\right) + \sum_{i=1}^{m_0} \frac{(\alpha-1)\phi(z_i)}{\Phi(z_i)}\left(-1 + z_i^2 + \frac{z_i\phi(z_i)}{\Phi(z_i)}\right) + 2\sum_{i=1}^{n_0} z_i\right], \\
-h_{13} = -h_{31} &= \frac{1}{\eta}\left[\sum_{j=1}^{k} \frac{m_j \phi(\gamma_j)}{\Phi(\gamma_j)} + \sum_{i=1}^{n_0} \frac{\phi(z_i)}{\Phi(z_i)}\right], \\
-h_{22} &= \frac{1}{\eta^2}\left[\sum_{j=1}^{k} \frac{\alpha m_j \gamma_j \phi(\gamma_j)}{\Phi(\gamma_j)}\left(-2 + \gamma_j^2 + \frac{\gamma_j\phi(\gamma_j)}{\Phi(\gamma_j)}\right) + \sum_{i=1}^{n_0} \frac{(\alpha-1)z_i\phi(z_i)}{\Phi(z_i)}\left(-2 + z_i^2 + \frac{z_i\phi(z_i)}{\Phi(z_i)}\right) + 3\sum_{i=1}^{n_0} z_i^2\right], \\
-h_{23} = -h_{32} &= \frac{1}{\eta}\left[\sum_{j=1}^{k} \frac{m_j \gamma_i \phi(\gamma_j)}{\Phi(\gamma_j)} + \sum_{i=1}^{n_0} \frac{z_i\phi(z_i)}{\Phi(z_i)}\right], \\
-h_{33} &= \frac{n_0}{\alpha^2}
\end{aligned}$$

Let $n = n_0 + \sum_{i=1}^{k} m_i$. As the Fisher information matrix is $\mathbf{I} = -E[\mathbf{H}(\xi,\eta,\alpha)]$. The observed Fisher information matrix is given by

$$\mathbf{I}\widehat{(\xi,\eta,\alpha)} = -\frac{1}{n}\begin{pmatrix} h_{11}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{12}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{13}(\hat{\xi},\hat{\eta},\hat{\alpha}) \\ h_{21}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{22}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{23}(\hat{\xi},\hat{\eta},\hat{\alpha}) \\ h_{31}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{32}(\hat{\xi},\hat{\eta},\hat{\alpha}) & h_{33}(\hat{\xi},\hat{\eta},\hat{\alpha}) \end{pmatrix} \tag{25}$$

The estimated covariance matrix of $(\hat{\xi}, \hat{\eta}, \hat{\alpha})$ is explicitly given by

$$V \equiv \operatorname{cov}\begin{pmatrix}\hat{\xi}\\ \hat{\eta}\\ \hat{\alpha}\end{pmatrix} = \mathbf{I}^{-1}(\xi, \eta, \alpha). \tag{26}$$

The standard large sample theory yields the following asymptotic multivariate normal distribution

$$(\hat{\xi} - \xi_0, \hat{\eta} - \eta_0, \hat{\alpha} - \alpha_0)^T \sim \mathbf{N}(\mathbf{0}, V) \tag{27}$$

where ξ_0, η_0 and α_0 are the true values of the population mean and the standard deviation.

4. Confidence Intervals of the Mean and Variance

4.1. Asymptotic Confidence Intervals

It is common in environmental and related areas that the measurements of interest are asymmetrically distributed. We are interested in making inference about the mean and variance of the underlying population. Recall that

$$\mu = E[X] = \int_0^\infty x d\left[\Phi\left(\frac{\log x - \xi}{\eta}\right)\right]^\alpha = \int_0^1 \alpha t^{\alpha-1} \exp\left[\xi + \eta\Phi^{-1}(t)\right] dt, \tag{28}$$

$$\sigma^2 = \int_0^\infty x^2 d\left[\Phi\left(\frac{\log x - \xi}{\eta}\right)\right]^\alpha - \mu^2 = \int_0^1 \alpha t^{\alpha-1} \exp\left[2\xi + 2\eta\Phi^{-1}(t)\right] dt - \mu^2. \tag{29}$$

Using plug-in principle of MLE, we have MLE of population mean and variance as follows

$$\hat{\mu} = \int_0^1 \hat{\alpha} t^{\hat{\alpha}-1} \exp\left[\hat{\xi} + \hat{\eta}\Phi^{-1}(t)\right] dt, \tag{30}$$

$$\hat{\sigma}^2 = \int_0^1 \hat{\alpha} t^{\hat{\alpha}-1} \exp\left[2\hat{\xi} + 2\hat{\eta}\Phi^{-1}(t)\right] dt - \hat{\mu}^2. \tag{31}$$

Next we establish asymptotic confidence intervals for both population mean μ and variance σ^2. To construct the confidence interval for the mean, we need following partial derivatives:

$$\begin{aligned}
\mu_\xi &= \frac{\partial\mu}{\partial\xi} = \int_0^1 \alpha t^{\alpha-1} \exp[\xi + \eta\Phi^{-1}(t)]dt,\\
\mu_\eta &= \frac{\partial\mu}{\partial\eta} = \int_0^1 \alpha t^{\alpha-1}\Phi^{-1}(t) \exp[\xi + \eta\Phi^{-1}(t)]dt,\\
\mu_\alpha &= \frac{\partial\mu}{\partial\alpha} = \int_0^1 (t + \alpha^2 - \alpha)t^{\alpha-2} \exp[\xi + \eta\Phi^{-1}(t)]dt,\\
\sigma_\xi^2 &= \frac{\partial\sigma^2}{\partial\xi} = 2\int_0^1 \alpha t^{\alpha-1} \exp[2\xi + 2\eta\Phi^{-1}(t)]dt,\\
\sigma_\eta^2 &= \frac{\partial\sigma^2}{\partial\eta} = 2\int_0^1 \alpha t^{\alpha-1}\Phi^{-1}(t) \exp[2\xi + 2\eta\Phi^{-1}(t)]dt,\\
\sigma_\alpha^2 &= \frac{\partial\sigma^2}{\partial\alpha} = \int_0^1 (t + \alpha^2 - \alpha)t^{\alpha-2} \exp[2\xi + 2\eta\Phi^{-1}(t)]dt - 2\mu\mu_\alpha.
\end{aligned}$$

Next we linearize the MLE of mean and variance by using the first order Taylor expansion at true values of (ξ, η, α), denoted by $(\xi_0, \eta_0, \alpha_0)$. Let $(\mu_{\xi_0}, \mu_{\eta_0}, \mu_{\alpha_0})$ and $(\sigma_{\xi_0}^2, \sigma_{\eta_0}^2, \sigma_{\alpha_0}^2)$ be the true value of $(\mu_\xi, \mu_\eta, \mu_\alpha)$ and $(\sigma_\xi^2, \sigma_\eta^2, \sigma_\alpha^2)$, respectively. The resulting Taylor expansion of $\hat{\mu}$ and $\hat{\sigma}^2$ are given by

$$\hat{\mu} = \mu_0 + \mu_{\xi_0}(\hat{\xi} - \xi_0) + \mu_{\eta_0}(\hat{\eta} - \eta_0) + \mu_{\alpha_0}(\hat{\alpha} - \alpha_0) + o_P(n^{-1/2}),$$

and

$$\hat{\sigma}^2 = \sigma_0^2 + \sigma_{\xi_0}^2(\hat{\xi} - \xi_0) + \sigma_{\eta_0}^2(\hat{\eta} - \eta_0) + \sigma_{\alpha_0}^2(\hat{\alpha} - \alpha_0) + o_P(n^{-1/2}).$$

Define

$$\mathcal{D}_0 = \begin{pmatrix} \mu_{\xi_0} & \mu_{\eta_0} & \mu_{\alpha_0} \\ \sigma_{\xi_0}^2 & \sigma_{\eta_0}^2 & \sigma_{\alpha_0}^2 \end{pmatrix}. \tag{32}$$

With the notations defined above, we summarize the asymptotic distribution of $(\hat{\mu}, \hat{\sigma}^2)^T$ in the following theorem.

Theorem 4.1. *When the sample size is large, we have*

$$(\hat{\mu} - \mu_0, \hat{\sigma}^2 - \sigma_0^2)^T \rightharpoonup \mathbb{N}_3(\mathbf{0}, \Sigma_0) \tag{33}$$

where $\Sigma_0 = \mathcal{D}_0 V_0 \mathcal{D}_0^T$, V *is specified in (26) and* $V_0 = \mathbf{I}^{-1}(\xi_0, \eta_0, \alpha_0)$.

In practical application, $\mathcal{D}_0$ is replaced by its MLE. That is, $\widehat{\mathcal{D}}_0$ is obtained by replacing $(\xi_0, \eta_0, \alpha_0)$ with $(\hat{\xi}, \hat{\eta}, \hat{\alpha})$. The MLE of Σ is given by $\hat{\Sigma} = \hat{\mathcal{D}}_0 \hat{V} \hat{\mathcal{D}}_0^T$, where V is specifed in (26)). Let $V_d = (v_{11}, v_{22}, v_{33})$ be the main diagonal elements of Σ and $\hat{V}_{11} = (\hat{v}_{11}, \hat{v}_{22}, \hat{v}_{33})$ be the main diagonal of $\hat{\Sigma}$.

Corollary 4.2. *When sample size is large, the* $100(1-\theta)\%$ *two sided confidence intervals for* μ *and* σ^2 *are given, respectively, by*

$$\hat{\mu} \pm Z_{1-\theta/2}\sqrt{\hat{v}_{11}} \quad \text{and} \quad \hat{\sigma}^2 \pm Z_{1-\theta/2}\sqrt{\hat{v}_{22}}$$

where $Z_{1-\theta/2}$ *is the normal critical value at confidence level* $100(1-\theta)\%$.

These confidence interval will be implemented with real-life data in a later section. Next we introduce the inference of process quantile estimation to be used in environmental process control and monitoring.

5. Inference on Quantiles and Applications

Estimation of quantile (also called value-at-risk) is also important in environmental process such as cleanliness, temperature, relative humidity, pressurization, vibration and acoustic isolation, exposure limits, etc. Let X_q be the q-th quantile we are interested in estimating (the tail probability q is usually large and close to 1). Using the CDF of ELN given in (1), we have

$$q = \left[\Phi\left(\frac{\log(X_q) - \xi}{\eta}\right)\right]^\alpha.$$

Solve X_q from the above equation, we obtain the q-th quantile as follows

$$X_q = \exp\left[\eta\Phi^{-1}(q^{1/\alpha}) + \xi\right] \tag{34}$$

Plugging the MLE of (ξ, η, α) into the above quantile, we have MLE of $\hat{X}_q$. To construct the confidence interval for X_q, need following partial derivatives.

$$\begin{aligned}
X_q^\xi &= \frac{\partial X_q}{\partial \xi} = \exp\left[\eta\Phi^{-1}(q^{1/\alpha}) + \xi\right], \\
X_q^\eta &= \frac{\partial X_q}{\partial \eta} = \exp\left[\eta\Phi^{-1}(q^{1/\alpha}) + \xi\right]\Phi^{-1}(q^{1/\alpha}), \\
X_q^\alpha &= \frac{\partial X_q}{\partial \sigma} = \exp\left[\eta\Phi^{-1}(q^{1/\alpha}) + \xi\right]\frac{\eta q^{1/\alpha}\log q}{\alpha^2\phi\left[\Phi(q^{1/\alpha})\right]}.
\end{aligned}$$

Let $D_{xq} = (X_q^\xi, X_q^\eta, X_q^\alpha)$. Let X_q^0 and D_{xq}^0 be the true value q-th quantile and vector D_{xq}^0, respectively. Then

$$\hat{X}_q = X_q + D_{xq}^0(\hat{\xi} - \xi_0, \hat{\eta} - \eta_0, \hat{\sigma}^2 - \sigma_0^2)^T + o_P(n^{-1/2}).$$

The asymptotic distribution of the MLE of $\hat{X}_q$ and confidence interval are given in the following theorem.

Theorem 5.1. *When the sample size is large, we have*

$$\hat{X}_q - X_q \to \mathbb{N}(0, D_{xq} V D_{xq}^T) \tag{35}$$

Using the MLE proposed before for V and D_{xq}, denoted by $\hat{V}$ and $\hat{D}_{xq}$, before, we can construct the $100(1-\theta)\%$ one-sided asymptotic confidence interval for q-th quantile X_q as follows

$$\hat{X}_q + Z_q \sqrt{\hat{D}_{xq} \hat{V} \hat{D}_{xq}^T} \tag{36}$$

where Z_q is the normal critical value at confidence level $100q\%$.

The lower quantile is rarely used in the environmental study. However, the confidence interval for lower quantile can be similarly constructed with form $\hat{X}_{1-\theta} - Z_{1-\theta}\sqrt{\hat{D}_{xq} \hat{V} \hat{D}_{xq}^T}$, where θ is close to 0.

6. Goodness-of-fit Tests

One practical we have to address is how the model fits well. For complete data, many testing procedures have been proposed in literature. There are also some testing also available for case of right random type II censored and left truncated data. In environmental study, the censoring data are observed due to accuracy and precision. That is censoring data occurred at fixed detection limits of different measuring devices. In this paper, we adopt Pearson χ^2 goodness-of-fit test to assess the model fit of the exponentiated lognormal distribution.

The Pearson χ^2 test defined based on the comparison of the observed frequencies and the expected (theoretical) frequencies on bins (data windows). The way of defining bins may be different. We suggest following algorithm to specify bins and find corresponding observed and expected frequencies to be used in Pearson χ^2 goodness-of-fit test.

Algorithm Let $d = \max\{d_1, d_2, \cdots, d_k\}$ be maximum censoring of the k limits in the sample. The number of observations that are bigger than d, denoted by m_d, is large.

(1) Let $x_{\max}$ be the maximum observation of the data. Partition data interval $[m_d, x_{\max}]$ into appropriately, say $m_d = X_{q_0} < X_{q_1} < X_{q_2} < X_{q_3} < \cdots < X_{q_P} < x_{\max}$, such that each sub-interval contains more than five observations.

(2) Use the MLE of (ξ, η, σ^2) obtained in Section 3 to generate $n = \sum_{i=0}^{k} m_i$ data values from $ELN(\hat{\xi}, \hat{\eta}, \hat{\alpha})$, denoted by $\{y_1, y_2, \cdots, y_n\}$.

(3) For $1 \le i \le n$, denote $O_0 = \#\{x_i < d_m\} + \sum_{i=1}^{k}$, $O_j = \#\{X_{q_j} \le x_i < X_{q_{j+1}}\}$ for $j = 0, 1, \cdots, P-1$ and $O_P = \#\{x_i > X_{q_P}\}$; denote also $E_0 = \#\{y_i < d_m\} + \sum_{i=1}^{k}$, $E_j = \#\{X_{q_j} \le y_i < X_{q_{j+1}}\}$ for $j = 0, 1, \cdots, P-1$ and $E_P = \#\{y_i > X_{q_P}\}$.

The above algorithm guarantees that the minimum frequency of each partition in the observed data is bigger than 5. If a frequency of a partition of the generated data is less than 5, one should combine partitions to make the combined frequency bigger than 5. Note also that the partitions are not of being equal width. Next we state the Pearson χ^2 goodness-of-fit test.

Theorem 6.1. *Assume that the number of observations bigger than the maximum detection limit. Using above algorithm to find the observed and expected frequencies O_j and E_j for $j = 0, 1, \cdots, P$, we have*

$$T = \sum_{j=0}^{P} \frac{(Q_j - E_j)^2}{E_j} \sim \chi_P^2 \tag{37}$$

under the null hypothesis that the exponentiated lognormal distribution fits the data well.

If the p-value of one-sided test is less than the significance level, one should reject the null hypothesis. Parsimony is one of the desired feature in modeling. Since lognormal distribution is a special case of exponentiated lognormal distribution, it is necessary to develop a procedure to detect potential over-fit. In other words, we can test $\mathrm{H}_0 : \alpha = 1$ versus $\mathrm{H}_a : \alpha \neq 1$. There different ways to testing this hypothesis. For example, one can equivalently test the area under ROC curve given in (5) to be 0.5. In this paper, we use the asymptotic distribution of $(\hat{\xi}, \hat{\eta}, \hat{\alpha})$ in 27 and proposed the following Wald χ^2 test.

Theorem 6.2. *Assume that the sample size is large and the underlying distribution is exponentiated lognormal. We have*

$$W = \frac{\alpha - 1}{\hat{v}_{33}} \sim \chi_1^2, \tag{38}$$

where $\hat{v}_{33}$ is the third diagonal element of the MLE of V given in (26).

If the observed p-value based on (38) is less than the given significance level, we reject the null hypothesis that the underlying distribution is exponentiated lognormal. Otherwise, we should assume lognormal distribution.

7. Bootstrap Procedure under Left-censored Sample

Focuses on the design of bootstrapping procedures for multiply left censored data. Comparing bootstrap CIs with the asymptotic CIs.

8. Simulation Studies

We will perform some simulation studies about the performance of confidence intervals and hypothesis test of the population proposed in previous sections based on left censored data in this section.

8.1. Random Number Generation

We can use the inverse CDF function method to derive the random numbers to generate samples from the exponentiated log-normal distribution. For a given random number from uniform population $U[0,1]$, the inverse of the CDF (1) that has the following form

$$X = \exp\left[\xi + \eta\Phi^{-1}(U^{1/\alpha})\right] \tag{39}$$

is the random number from the exponentiated log-normal population of (1).

8.2. Small Sample Performance of MLE

9. Real-world Applications

Two numerical examples based on real-world applications in environmental studies.
In this section, we work on several real-world environmental data sets that contain single and multiple detection limits. We will report the MLE of parameters and various confidence intervals (asymptotic and bootstrap percentile and BCa). The precision used in the numeric procedure is 10^{-8} (the absolute difference between the true maximum likelihood and the likelihood evaluated at the estimated parameters is less than or equal to 10^{-8}). We use the gradient search method to optimize the likelihood kernel.

9.1. Water Quality

Total recoverable chromium concentrations in streamflows of Gales Creek in Oregon, USA (publicaly available at: https://waterdata.usgs.gov/monitoring-location/453026123063401/). The objective is to relate chromium concentrations including censored values to mean daily flows over time and by season (wet versus dry seasons). Two detection limits decreasing over time.

9.2. Thiamethoxam Concentrations in Pollen

The Ministry of the Environment, Conservation and Parks, with assistance from the Ontario Ministry of Agriculture, Food and Rural Affairs, started the Pollen Monitoring Network in 2015, in support of Ontario's Pollinator Health Action Plan. This was to measure changes in pesticide (neonicotinoid insecticide, insecticide and fungicide) concentrations following restrictions on the use of neonicotinoid-treated corn and soybean seeds. A publicaly available data set contains measurements of pesticide (neonicotinoid insecticide, insecticide and fungicide) concentrations in pollen samples collected from honeybee yards in Ontario. The raw data can be found at *https://data.ontario.ca/en/dataset/pollen-monitoring-network-study*. In this example, we only interested in the concentration of Thiamethoxam. A modified and reorganized version of this subset was built in the R package *NADA2*.

9.3. San Joaquin Valley Underground Water

7.1. Left Doubly Censored Sample Case

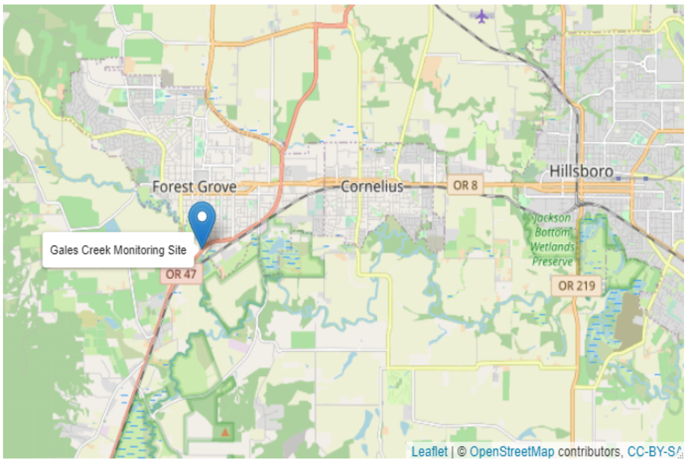

Figure 4.

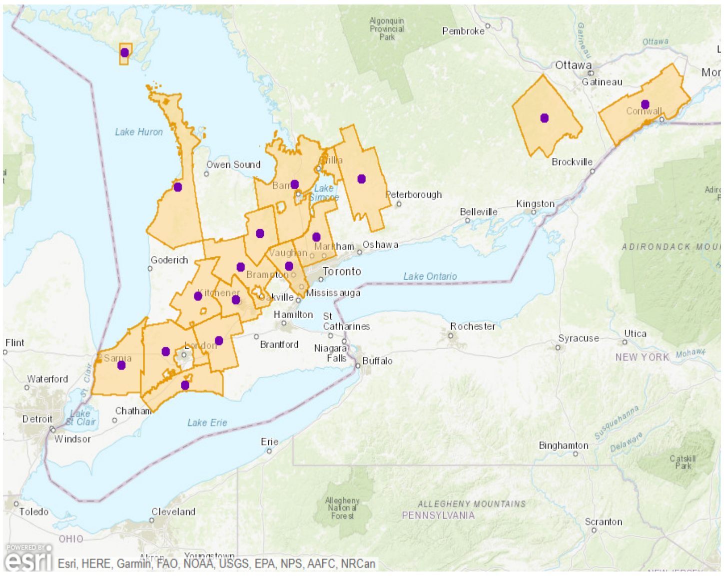

Figure 5. .

Example 7.1. **Ground Water Data** [6] studied the distribution of tracement concentrations in shallow ground-waters from Alluvial Fan Zone and Basin-Trough Zone underneath the San Joaquin Valley, California. The dataset contains multiple detection limits and a few missing values as well. [26] re-analyzed copper and zinc concentrations using nonparametric rank based two sample location tests to compare the copper and zinc concentrations in the two geologic zones. In this example, we use the 68 Zinc concentrations (including one missing value, denoted by *ms*) given in micrograms per liter: <10, 9, ms, 5, 18, <10, 12, 10, 11, 11, 19, 8, <3, <10, <10, 10, 10, 10, 10, <10, 10, <10, 10, <10, 10, <10, 10, 10, 20, 20, <10, 20, 20, 20, <10, 10, 20, 620, 40, 50, 33, 10, 20, 10, 10, 10, 30, 20, 10, 20, 20, 20, <10, 20, 23, 17, 10, <10, 10, 20, 29, 20, <10, 10, <10, 10, 7, <10.

In this analysis and thereafter, we will ignore the missing measurements (treat it as if it is missing at completely random, MCR). As point in [30] the original measurements are not normally distributed, we fit this data to exponentiated lognormal distribution $ELN(\xi, \eta, \alpha)$ to this data.

10. Concluding Remarks

We have extended the study of exponentiated lognormal distribution with environmental application by [25]. We studied some further properties of the shape parameter α theoretically. Theory and graphics show that adding a shape parameter α makes the distribution more flexible in terms of mean and variance. This paper focuses on fitting the data to left censored environmental data with detection limits and the inference of population mean and variance.

The likelihood principle was used in this study. To make the inferential procedure statistically rigorous, we proposed a Pearson χ^2 for the goodness-of-fit of the exponentiated lognormal to the left censored data and Wald χ^2 test for significance of the shape parameter to avoid potential over-fitting problem.

Algorithm 1 Merge Sort

```
function Merge(A, p, q, r)                ▷ Where A - array, p - left, q - middle, r - right
    n_1 = q − p + 1
    n_2 = r − q
    Let L[1 . . . n_1 + 1] and R[1 . . . n_2 + 1] be new arrays
    for i = 1 to n_1 do
        L[i] = A[p + i − 1]
    end for
    for j = 1 to n_2 do
        R[i] = A[q + j]
    end for
    L[n_1 + 1] = ∞
    R[n_2 + 1] = ∞
    i = 1
    j = 1
    for k = p to r do
        if L[i] < R[j] then
            A[k] = L[i]
            i = i + 1
        else if L[i] > R[j] then
            A[k] = R[j]
            j = j + 1
        else
            A[k] = −∞                     ▷ We mark the duplicates with the largest negative integer
            j = j + 1
        end if
    end for
end function
```

The confidence interval of population mean and variances are based on the large sample distribution of the MLE of the mean and variance. We also provide an asymptotic confidence interval for population quantile (value-at-risk) of the exponentiated lognormal distribution.

The Pearson χ^2 goodness-of-fit test proposed in this article assume that the number of observations bigger than the maximum detection limit is large. In some practical situation, the assumption may not be satisfied. A more robust and powerful bootstrap goodness-of-fit of parametric families based on type I left censored samples will be studied in a separate study.

Since exponentiated lognormal distribution is a lifetime distribution, the methods proposed in this paper can be easily adjusted to survival analysis with right censored data.

References

[1] Cohen AC. 1957. On the solution of estimating equations for truncated and censored samples from normal populations, *Biometrika* **44**, 225–236.

[2] Cohen AC. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics* **1**, 217–237.

[3] Cohen AC. 1991. *Truncated and censored samples: Theory and applications*, Vol. 119 of *Statistics: Textbooks and Monographs*, Marcel Dekker Inc., New York.

[4] Davison AC. and Hinkley DV. 1997. *Bootstrap methods and their application*, Vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.

[5] Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B* **39**(1), 1–38.

[6] Deverel SJ, Millard SP. 1988. Disitribution and mobility of selenium and other trace elements in shallow ground water of the western San Joaquin Valley, California, *Environ. Sci. Technol.* **22**, 697–702.

[7] Efron B. 1979. Bootstrap methods: another look at the jackknife, *Ann. Statist.* **7**(1), 1–26.

[8] Efron B. 1981a. Censored data and the bootstrap, *J. Amer. Statist. Assoc.* **76**(374), 312–319.

[9] Efron B. 1981b. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods, *Biometrika* **68**(3), 589–599.

[10] Efron B. 1985. Bootstrap confidence intervals for a class of parametric problems, *Biometrika* **72**(1), 45–58.

[11] Efron B, Tibshirani RJ. 1993. *An introduction to the bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

[12] El-Shaarawi AH. 1989. Inferences about the mean from censored water quality data, *Water Resour. Res.* **25**(4), 685–690.

[13] El-Shaarawi AH, Esterby SR. 1992. Replacement of censored observations by a constant: an evaluation, *Water Res.* **26**(6), 835–844.

[14] Fletcher R. 1970. A new approach to variable metric algorithms, *The Computer Journal* **13**(3), 317–322.

[15] Fletcher R, Reeve CM. 1964. Function minimization by conjugate gradients, *The Computer Journal* **7**(2), 149–154.

[16] Gilliom R, Helsel D. 1986. Estimation of distributional parameters for censored trace level water quality data 1. estimation techniques, *Water Resour. Res.* **22**(2), 135–146.

[17] Goldfarb D. 1970. A family of variable-metric methods derived by variational means, *Math. Comp.* **24**, 23–26.

[18] Gupta AK. 1952. Estimation of the mean and standard deviat ion of a normal population from a censored sample, *Biometrika* **39**, 260–273.

[19] Martinez-Floreza, G, Bolfarineb H, and Gomez HW. 2014. The log-power-normal distribution with application to air pollution, *Environmetrics* **25**, 44–56.

[20] Helsel D. 1990. Less than obvious: statistical treatment of data below the detection limit, *Envion. Sci. Technol.* **24**(12), 1767–1774.

[21] Helsel D. 2005. *Nondetects and data analysis: statistics for censored environmetal data*, Wiley, New York.

[22] Helsel D, Gilliom R. 1986. Estimation of distributional parameters for censored trace level water quality data 2. verification and applications, *Water Resour. Res.* **22**(2), 147–155.

[23] Helsel D. 2006. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it, *Chemosphere,* **65**, 2434–2439.

[24] Marshall, AW, Olkin I. 2007. *Life Distributions - Structure of Nonparametric, Semiparametric, and Parametric Families*, Springer, NY, New York, USA.

[25] Martínez-Flórez, G, Bonfarine, H. and Gómez, HW. 2014. The log-power-normal distribution with application to air pollution, *Environmetrics*, **25**, 44-56.

[26] Millard SP, Deverel SJ. 1988. Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits, *Water Resour. Res.* **24**(12), 2087–2098.

[27] Pepe MS. 2004. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Inc. New York.

[28] Nelder JA, Mead R. 1965. A simplex method for function minimization, *The Computer Journal* **7**(4), 308–313.

[29] Persson T, Rootzen H. 1977. Simple and highly efficient estimators for a type I censored normal sample, *Biometrika* **64**, 123–128.

[30] Peng C. 2010. Interval estimation of population parameters based on environmental data with detection limits, *Environmetrics*, **21**(6), 645–658.

[31] Shanno DF. 1970. Conditioning of quasi-Newton methods for function minimization, *Math. Comp.* **24**, 647–656.

[32] Shumway RH, Azari AS, Johnson P. 1989. Estimating mean concentration under transformation for environmetal data with detection limits, *Technometrics* **31**, 347–356.

[33] Singh A, Nocerino J. 2002. Robust estimation of mean and variance using environmental data sets with below detection limit observations, *Chemometrics and Intelligent Laboratory Systems* **60**(1), 69–86.

[34] U.S. EPA. 1992. *Statistical Analysis of Ground Water Monitoring Data at RARA Facilities, Addendum to Interim Final Guidance*: Office of Solid Waste, Waste Management Division. U.S. EPA.

[35] Tiku ML. 1967. Estimating the mean and standard deviation from a censored normal sample, *Biometrika* **54**, 155–165.

[36] Zhao Y and Frey, HC. 2006. Uncertainty for data with non-detects: air toxic emissions from combustion, *Human and Ecological Risk Assessment.* **12**, 1171–1191.