Journal of Environmental Statistics

September 2018, Volume 8, Issue 7. http://www.jenvstat.org

Prediction Limits for the Mean of a Sample from a Lognormal Distribution: Uncensored and Censored Cases

K. Krishnamoorthy
Department of Mathematics
University of Louisiana at Lafayette

Md Sazib Hasan
Department of Mathematics
University of Louisiana at Lafayette

Abstract

For some regulatory purposes, it is desired to compare average on-site pollution concentrations in a narrowly defined geographic area with a large collection of background measurements. An approach to this problem is to treat this as a statistical prediction for the mean of a future sample based on a background sample. In this article, assuming lognormality, a fiducial approach is described for constructing prediction limits for the mean of a sample when the background sample is uncensored or censored. The fiducial prediction limits are evaluated with respect to coverage probabilities, and are compared with those based on another approximate method. Monte Carlo simulation studies for the uncensored case indicate that the fiducial methods are accurate and practically exact even for small samples, and they are very satisfactory for the censored case. Algorithms for computation of confidence limits are provided. The methods are illustrated using two real data sets.

Keywords: coverage probability, detection limits, fiducial quantity, maximum likelihood estimates, normal-based methods, prediction limits.

1. Introduction

The applicability of a lognormal distribution has been validated for several practical situations where the data are positive and right-skewed. Lognormal model was used for analyzing data on workplace exposure to contaminants by Oldham (1953), Esmen and Hammad (1977), Selvin and Rappaport (1989), and Lyles and Kupper (1996)), and Krishnamoorthy and Mathew (2003). The lognormal distribution has become a common choice to represent intrinsically positive and often highly skewed environmental data in statistical analysis; see Georgopoulos and Seinfeld (1982), Speer, and Waite (1975), Gilbert (1987) and Bhaumik and Gibbons (2004). Confidence limits (CLs) based on a sample for the mean of a lognormal distribution are used in environmental compliance and workplace pollution concerns. In particular, an upper confidence limit (UCL) for the mean of a lognormal distribution based on a sample is used to test if the pollution levels at some location are in compliance with regulatory standards. Land (1971) has proposed an exact method of finding confidence limits (CLs) for the mean, and others (e.g., Krishnamoorthy and Mathew, 2003 and Zou et al., 2009) have proposed simple accurate approximate methods. Solutions to other problems such as prediction limit for a single future observation, tolerance limits and confidence limits for an exceedance probability can be obtained in a straightforward manner using one-to-one relation between the lognormal and normal distributions.

Another important problem in environmental statistics, noted by Bhaumik and Gibbons (2004), is the comparison of the average of a small number of on-site measurements with a larger collection of background measurements. Bhaumik and Gibbons have described an example in which a series of on-site soil samples collected in an area of potential environmental concern are compared with the background concentration distribution, characterized by n background measurements. Bhaumik and Gibbons (2004) treated this problem as statistical prediction for the mean of a future sample based on a background sample. Applying the normal-based method to log-transformed sample, one could easily obtain a prediction limit for the geometric mean of a future sample from the same lognormal distribution. However, such normal-based methods can not be applied directly to find a prediction limit for the arithmetic mean of a future sample from a lognormal distribution.

A hindrance in industrial hygiene and environmental data analysis is dealing with samples that include some concentration levels below detection limits (DLs), thus resulting in nondetect values. Samples with multiple DLs arise when the measurements are obtained using different devices, each with its own limitation of detecting contaminant levels, or samples are analyzed by different laboratories. Early recommendations were to replace the nondetects with a fraction of DL, but such substitution methods often lead to inaccurate results; see the editorial note by Ogden (2010). So the goal of this article is to propose a simple method to find a prediction limit for a future sample mean, and the method that can be easily extended to samples with multiple detection limits. Towards this goal, we propose the fiducial approach (Fisher, 1935) to log-transformed samples. The fiducial approach, in the name of *generalized variable approach* (Tsui and Weerahandi, 1989), has been used to find confidence limits for the mean of a lognormal distribution and to find a confidence interval (CI) for the difference between or the ratio of means of two lognormal distributions by Krishnamoorthy and Mathew (2003). The so-called generalized variable approach, introduced by Tsui and Weerahandi (1989) and Weerahandi (1993), is a special case of the fiducial inference introduced by Fisher (1930, 1935); see Hannig, Iyer and Patterson (2006). For the continuous case, the fiducial approach has been used successfully to estimate or to test a function of parameters where pivotal quantities are available for individual parameters (e.g., lognormal mean, normal quantiles and quantiles in one-way random model). Wang, Hannig and Iyer (2012) have proposed a fiducial approach to find prediction intervals (PIs) for the mean of a future sample in a general setup, and illustrated their approach for finding prediction limits for the cases of normal, gamma, exponential and Weibull distributions. These authors have noted that fiducial prediction limits for various random quantities are comparable with those based on other exact methods (e.g., normal, exponential and Weibull). Recently, Krishnamoorthy and Wang (2016) have applied the fiducial approach to cube root transformed samples from a gamma distribution to obtain prediction limits for the mean of a future sample. Their results can be used in a straightforward manner to obtain prediction limits when the samples include detection limits.

It should be noted that a PI for a single future observation from a lognormal distribution can be obtained in a straightforward manner (see Section 2.3), but finding PI for the mean of a future sample is not a trivial task. The paper by Bhaumik and Gibbons (2004) appears to be the first one addressing this problem. As will be seen in the sequel of this paper, this approach is quite complex, not satisfactory in terms of coverage probability, and is not easy to extend to the case of multiple detection limits. Recently, Martin and Lingham (2016) have provided a general approach referred to as "prior-free probabilistic prediction" based on inferential modelling (IM) and illustrated the approach for finding a prediction interval in our present problem. It is well-known that a fiducial distribution for a parameter is a prior-free posterior distribution (Efron, 1998), and so it is natural to expect that the fiducial PI should be similar to the one in Martin and Lingham (2016). Indeed, the method that we shall employ is based on the fiducial approach by Wang et al. (2012), and the resulting PI is identical to the one given in Martin and Lingam (2016). Furthermore, fiducial approach can be readily extended to the censored case.

The rest of the article is organized as follows. In the following section, we outline the method by Bhaumik and Gibbons (2004) for finding an upper prediction limit for the mean of a future sample. We also describe fiducial quantities (FQs) for μ and σ^2, which are functions of the mean and variance of log-transformed samples and some random variables whose distributions do not depend on any unknown parameters. On the basis of the fiducial quantities, and following the approach by Wang et al. (2012), we outline a simulation-based approach to find a prediction limit for the mean of a future sample. The fiducial inference, in general, is not exact. For the case of predicting a single future observation, we show that the fiducial PI is exact in the frequentist sense, and for the case of predicting the mean of future sample, our simulation studies indicate the fiducial PIs are very accurate. The results are extended to samples with multiple detection limits in Section 3. The coverage probabilities of the proposed prediction limits were evaluated using Monte Carlo simulation, and they were compared with those of the prediction limits by Bhaumik and Gibbons (2004). In Section 4, the fiducial methods are illustrated using two environmental data sets, one is uncensored and another includes two detection limits. Some concluding remarks are given in Section 5.

2. Uncensored Case

2.1. Bhaumik-Gibbons' Approach

To outline the approach by Bhaumik and Gibbons (2004), let $Y_1, ..., Y_n$ be a background sample and $Y_1^*, ..., Y_{n_*}^*$ be a future sample from a lognormal distribution with parameters $E(\ln Y_i) = \mu$ and $\text{var}(\ln Y_i) = \sigma^2$. Let T^* denote the total of the future sample. Bhaumik and Gibbons have used an approximate probability density function (pdf) of a studentized version Z^* of T^* to find an upper prediction limit (UPL) for a future sample mean. If $f(z^*|\mu, \sigma)$ denote the approximate pdf of Z^*, then a $100(1-\alpha)\%$ UPL for the mean of the future sample can be obtained from T_U, where T_U is determined by the equation

$$\int_0^{T_U} f(z^*|\widehat{\mu},\widehat{\sigma})dz^* = 1-\alpha.$$

Bhaumik and Gibbons have provided an explicit expression for $f(z^*|\widehat{\mu},\widehat{\sigma})$, which does not even satisfy some basic requirements of a pdf; it could be negative with multiple modes over a range of Z^* and the integral over the support of Z^* could be different from unity. Furthermore, the pdf used in the above equation is estimated by replacing the unknown parameters with the estimates $\widehat{\mu}$ and $\widehat{\sigma}$, and so the predicting method does not account for the variability of the background samples. For these reasons, any method on the basis of an approximate pdf of Z^* could lead to inaccurate results. In fact, to find a prediction limit for $\bar{Y}^*$, one needs the pdf of a studentized quantity $(\bar{Y}-\bar{Y}^*)/(\widehat{\text{var}}(\bar{Y}-\bar{Y}^*))$, and finding such pdf is not an easy task.

2.2. Fiducial Quantities

In order to apply the fiducial approach by Wang et al. (2012) to find a UPL for a future sample mean, we first need to describe the fiducial quantities for the lognormal parameters. Consider a sample $Y_1,...,Y_n$ from a lognormal distribution with parameters μ and σ^2. Let $X_i = \ln(Y_i)$, $i = 1,...,n$, so that X_i's are independent $N(\mu,\sigma^2)$ random variables. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2. \tag{1}$$

To describe the fiducial quantities for μ and σ^2 on the basis of Dawid and Stone (1982) approach, we first note that

$$\bar{X} \overset{d}{=} \mu + Z\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad S^2 \overset{d}{=} \sigma^2\frac{\chi^2_{n-1}}{n-1},$$

where Z is the standard normal random variable, χ^2_m denotes the chi-square random variable with degrees of freedom (df) m and the notation "$\overset{d}{=}$" means "distributed as." Furthermore, Z and χ^2_{n-1} are independent random variables. Let $(\bar{x},s)$ be an observed value of $(\bar{X},S)$. Solving the above equations for μ and σ^2, and replacing $(\bar{X},S)$ with $(\bar{x},s)$, we obtain the fiducial quantities for the parameters as

$$G_\mu = \bar{x} + \frac{Z\sqrt{n-1}}{\sqrt{\chi^2_{n-1}}}\frac{s}{\sqrt{n}} \quad \text{and} \quad G_{\sigma^2} = \frac{(n-1)s^2}{\chi^2_{n-1}}. \tag{2}$$

Notice that the fiducial distributions (conditional distributions of G_μ and G_{σ^2} given $(\bar{x},s)$) for the parameters do not depend on any unknown parameters. A fiducial quantity for a real-valued function, say, $h(\mu,\sigma^2)$ can be obtained by simple substitution as $h\left(G_\mu, G_{\sigma^2}\right)$. For example, a fiducial quantity for the mean of a lognormal distribution is given by $G_M = \exp\left(G_\mu + G_{\sigma^2}/2\right)$. For a given $\bar{x}$ and s^2, the distribution of G_M does not depend on any unknown parameters, so its percentiles may be estimated by Monte Carlo simulation. More details, on finding the fiducial CIs for the mean and difference between two means, see Krishnamoorthy and Mathew (2003).

2.3. Fiducial Prediction Limits

Let $Y_1, ..., Y_{n_*}$ be a future sample from the lognormal distribution from which the background sample was obtained. Then

$$(Y_1, ..., Y_{n_*}) \overset{d}{=} \left(e^{\mu+Z_1^*\sigma}, ..., e^{\mu+Z_{n_*}^*\sigma}\right), \tag{3}$$

where Z_i^*'s are independent standard normal random variables. Let $(\bar{x}, s^2)$ be the observed value of (mean, var) based on a log-transformed background sample of size n. Replacing μ and σ in (3) with the fiducial quantities G_μ and G_σ in (2), we obtain fiducial variables for a future sample as

$$(G_{Y_1}, ..., G_{Y_{n_*}}) = \left(e^{G_\mu+Z_1^*G_\sigma}, ..., e^{G_\mu+Z_{n^*}^*G_\sigma}\right). \tag{4}$$

The above fiducial variables are a function of the standard normal random variables and the fiducial variables G_μ and G_σ, whose distributions do not depend on any parameters. So, for a given $(\bar{x}, s)$, the joint distribution can be estimated by Monte Carlo simulation. In particular, for a given $(\bar{x}, s)$, the percentiles of

$$\overline{G}_Y = \frac{1}{n_*}\sum_{j=1}^{n_*} G_{Y_j} \tag{5}$$

can be estimated by Monte Carlo simulation. The $100(1-\alpha)$ percentile of $\overline{G}_Y$ is the fiducial UPL for the mean of a future sample; for more details, see Algorithm 1. We observe from (4) that the sampling distribution of $\overline{G}_Y$ is determined by the future sample size n^* and the background sample via the fiducial quantities G_μ and G_σ.

To shed some light on the fiducial prediction limit, let us consider the case of $n_* = 1$. That is, prediction limit for a single future observation Y from a lognormal distribution based on a background sample of size n. The UPL for Y, on the basis of normal-based prediction limit of $\ln Y$, is given by

$$\exp\left(\bar{x} + t_{n-1;1-\alpha}s\sqrt{1+\frac{1}{n}}\right), \tag{6}$$

where $t_{f;q}$ denotes the $100q$ percentile of the t distribution with df $= f$. To simplify the fiducial UPL based on (4) for the case of $n_* = 1$, let Z_1 and Z_2 be independent standard normal random variables. Using the expressions for G_μ and G_σ in (2), we see that

$$\begin{aligned} G_Y &\overset{d}{=} \exp\left(G_\mu + Z_2G_\sigma\right) \\ &= \exp\left(\bar{x} + s\sqrt{n-1}\frac{Z_1/\sqrt{n}+Z_2}{\sqrt{\chi^2_{n-1}}}\right) \\ &\overset{d}{=} \exp\left(\bar{x} + t_{n-1}s\sqrt{1+\frac{1}{n}}\right). \end{aligned}$$

To arrive at the third step, we used the facts that $Z_1/\sqrt{n} + Z_2$ is distributed as $\sqrt{1+1/n}$ times a standard normal random variable, and $\sqrt{m}Z/\sqrt{\chi^2_m} \sim t_m$. Thus, on the basis of the above stochastic representation, we see that the fiducial UPL for Y coincides with the exact one in (6).

To develop an algorithm for computing the prediction limit for the case of $n^* \geq 2$, we first note that G_{Y_j} in (4) has the stochastic representation that

$$\begin{aligned} \ln(G_{Y_j}) &\overset{d}{=} G_\mu + Z_j G_\sigma \\ &= \bar{x} + s\sqrt{n-1}\frac{Z/\sqrt{n} + Z_j}{\sqrt{\chi^2_{n-1}}}, \; j = 1, ..., n_*. \end{aligned}$$

Thus,

$$\overline{G}_Y = \exp(\bar{x}) \left(\frac{1}{n_*} \sum_{j=1}^{n_*} \exp(W_j) \right), \qquad (7)$$

where $W_j = s\sqrt{n-1}\frac{Z/\sqrt{n}+Z_j}{\sqrt{\chi^2_{n-1}}}$, $j = 1, ..., n_*$. The conditional distribution of $\overline{G}_Y$ given $(\bar{x}, s)$ is referred to as the fiducial predicting distribution. As noted in the introduction Martin and Lingham (2016) have obtained the same expression for $\overline{G}_Y$ using inferential modeling, and they refer to the conditional distribution of $\overline{G}_Y$ as the "plausible predicting distribution."

Using the expression (7) for $\overline{G}_Y$, calculation of the UPL can be carried out as shown in the following algorithm.

Algorithm 1

1. For a given sample from a lognormal distribution, compute the mean $\bar{x}$ and variance s^2 of the log-transformed sample.
2. Generate U from χ^2_{n-1} distribution and a normal variate Z from $N(0,1)$ distribution.
3. Generate a set of random numbers $Z_1, ..., Z_{n_*}$ from $N(0,1)$, and set
$$W_j = s\sqrt{n-1}(Z/\sqrt{n} + Z_j)/\sqrt{U}, \; j = 1, ..., n_*.$$
4. Compute $\overline{G}_Y = \exp(\bar{x}) \left(\frac{1}{n_*} \sum_{j=1}^{n_*} \exp(W_j) \right)$.
5. Repeat steps 2 – 4 for a large number of times, say, 100,000
6. The $100(1-\alpha)$ percentile of these 100,000 $\overline{G}_Y$'s is a $100(1-\alpha)\%$ UPL for the mean of a future sample of size n_*.

The above algorithm can be easily coded in any programming language. R codes of the algorithm is posted at www.ucs.louisiana.edu/∼kxk4695/ and is also available as a supplementary file at the journal's website.

2.4. Coverage Studies

We have shown in Section 2.3 that the fiducial UPL is exact for predicting a single future observation. However, the accuracy of the fiducial method should be judged by Monte Carlo simulation studies when the size of a future sample is two or more. Accordingly, we evaluate the coverage probabilities of the fiducial prediction limits for the sample size and parameter configurations considered in Bhaumik and Gibbons (2004) so that the fiducial method can be compared with the approximate method by these authors. The coverage probabilities of the fiducial method for the uncensored case are estimated as follows. We first generated 10,000 pairs of background and future samples of sizes n and n_*, respectively, from an assumed lognormal distribution. For each generated background sample, we estimated the 95% UPL for the mean of a future sample using Algorithm 1 with 10,000 runs. The proportion of the UPLs that include the means of the corresponding future samples is an estimate of the coverage probability.

The estimated coverage probabilities of 95% fiducial UPLs along with those of Bhaumik and Gibbons' (B-G) approach are reported in Table 1 for the uncensored case. The coverage probabilities of B-G UPLs are close to the nominal level when σ^2 is .0625 or .2. Even for such small values of σ^2, the B-G UPLs could be conservative for moderate to large samples. For $\sigma^2 = 2$ or 3, the B-G UPLs are anti-conservative for small n and conservative for large n. For example, when $(n, n_*, \sigma^2) = (5, 10, 3)$, the coverage probability of the B-G UPL is .859, and when $(n, n_*, \sigma^2) = (100, 10, 3)$, it is .977. On the other hand, the coverage probabilities of the fiducial UPLs were close to the nominal level for all the combinations of sample sizes and parameter values that we considered.

To judge the accuracy of the fiducial UPLs for very small sample sizes, we estimated the coverage probabilities for different confidence levels. The estimated coverage probabilities reported in Table 2 are practically coincide with the nominal confidence levels in all cases. This suggests that fiducial prediction intervals are satisfactory regardless of sample sizes.

Table 1: Coverage probabilities of 95% UPLs for the mean of a future sample of size n_* based on a background sample of size n

	$\mu = 2, \sigma^2 = .0625$				$\mu = 3, \sigma^2 = .2$			
	$n_* = 5$		$n_* = 10$		$n_* = 5$		$n_* = 10$	
n	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial
5	.949	.953	.949	.951	.935	.953	.949	.951
10	.976	.948	.972	.953	.966	.954	.972	.953
20	.981	.946	.981	.952	.974	.948	.981	.952
30	.984	.947	.984	.953	.978	.951	.984	.953
100	.987	.953	.989	.946	.982	.953	.989	.946
	$\mu = 3, \sigma^2 = .5$				$\mu = 3, \sigma^2 = 1$			
n	$n_* = 5$		$n_* = 10$		$n_* = 5$		$n_* = 10$	
n	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial
5	.917	.951	.910	.952	.935	.953	.887	.953
10	.955	.945	.943	.948	.966	.954	.931	.949
20	.965	.954	.959	.952	.974	.948	.950	.951
30	.971	.952	.966	.944	.978	.951	.958	.951
100	.977	.952	.975	.948	.982	.953	.972	.947
	$\mu = 3, \sigma^2 = 2$				$\mu = 3, \sigma^2 = 3$			
n	$n_* = 5$		$n_* = 10$		$n_* = 5$		$n_* = 10$	
n	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial	B-G	Fiducial
5	.890	.948	.867	.954	.887	.949	.859	.952
10	.942	.949	.920	.949	.943	.957	.916	.955
20	.959	.950	.950	.947	.963	.948	.953	.948
30	.968	.950	.960	.952	.973	.950	.963	.950
100	.977	.948	.976	.946	.978	.956	.977	.951

NOTE: B-G: Bhaumik and Gibbons' (2004) method; numbers in parentheses are expected values of UPLs.

Table 2: Coverage probabilities of 95% fiducial UPLs for small sample sizes

$\mu = 1$

σ	$(n, n_*) = (3, 2)$			$(n, n_*) = (4, 3)$			$(n, n_*) = (5, 3)$			$(n, n_*) = (6, 4)$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%
.2	.900	.950	.991	.904	.951	.991	.904	.951	.989	.907	.956	.993
.5	.909	.955	.992	.902	.952	.991	.907	.954	.990	.906	.952	.989
1	.907	.953	.991	.900	.952	.991	.906	.953	.990	.909	.954	.990
2	.900	.948	.988	.907	.953	.992	.901	.952	.991	.898	.947	.989
3	.903	.953	.990	.900	.948	.990	.903	.951	.990	.899	.950	.989

3. Prediction Limits: Censored Case

For the uncensored case, the fiducial quantities are a function of the mean and standard deviation based on log-transformed samples. We shall develop similar fiducial quantities as a function of the maximum likelihood estimates (MLEs) based on log-transformed censored background sample and other random quantities whose distributions do not depend on any unknown parameters. Such fiducial quantities are already obtained in Krishnamoorthy and Xu (2011), and later used in Krishnamoorthy, Mathew and Xu (2014) and Krishnamoorthy and Wang (2016) to obtain approximate solutions to some other problems involving normal and gamma distributions. For the sake of completeness and ease of reference, we shall describe the approach below.

3.1. Maximum Likelihood Estimates and Fiducial Quantities

Consider a simple random sample of n observations subject to k detection limits, say, DL_1,...,DL_k, from a normal distribution with mean μ and variance σ^2. Let us assume without loss of generality that $DL_1 < DL_2 < ... < DL_k$, and all the measurements are expressed in the same measurement unit. Let m_i denotes the number of nondetects that are below DL_i, and let $m = \sum_{i=1}^{k} m_i$, so that the number of detected observations is $n - m$. Let us denote the detected observations by $x_1, ..., x_{n-m}$. Define

$$\bar{x}_d = \frac{1}{n-m}\sum_{i=1}^{n-m} x_i \quad \text{and} \quad s_d^2 = \frac{1}{n-m}\sum_{i=1}^{n-m}(x_i - \bar{x}_d)^2. \tag{8}$$

Notice that $(\bar{x}_d, s_d^2)$ is the (mean, variance) based on the detected observations. The log-likelihood function for the censored case, after omitting a constant term, can be written as

$$l(\mu, \sigma) = \sum_{i=1}^{k} m_i \ln \Phi(z_i^*) - (n-m)\ln\sigma - \frac{(n-m)(s_d^2 + (\bar{x}_d - \mu)^2)}{2\sigma^2}, \tag{9}$$

where $z_i^* = \frac{DL_i - \mu}{\sigma}$, $i = 1, ..., k$. The MLE $(\widehat{\mu}, \widehat{\sigma})$ of (μ, σ) is obtained by maximizing (9) (Krishnamoorthy, Mathew and Xu, 2014).

In order to obtain fiducial quantities for μ and σ, we shall use the approximate distributional results for the MLEs given in Krishnamoorthy and Xu (2011). Let $(\widehat{\mu}, \widehat{\sigma})$ denote the MLE of (μ, σ) based on a log-transformed sample. Let $\widehat{\mu}^*$ and $\widehat{\sigma}^*$ denote the MLEs based on a sample of size n from a $N(0,1)$ distribution with DLs $DL_i^* = (\ln DL_i - \widehat{\mu})/\widehat{\sigma}$, $i = 1, ..., k$. Then

$$\frac{\widehat{\mu}-\mu}{\sigma} \sim \widehat{\mu}^* \quad \text{and} \quad \frac{\widehat{\sigma}}{\sigma} \sim \widehat{\sigma}^*, \text{ approximately.} \tag{10}$$

The above distributional results are exact if $P_i = \Phi\left((DL_i - \mu)/\sigma\right)$, $i = 1, ..., k$ are known (Krishnamoorthy and Zou, 2011). On the basis of the above results, we have the following approximate stochastic representations:

$$\widehat{\mu} \overset{d}{=} \mu + \widehat{\mu}^*\sigma \quad \text{and} \quad \widehat{\sigma} \overset{d}{=} \sigma\widehat{\sigma}^*, \text{ approximately.}$$

Let $(\widehat{\mu}_0, \widehat{\sigma}_0)$ be an observed value of $(\widehat{\mu}, \widehat{\sigma})$. Solving the above equations for μ and σ, and then replacing $(\widehat{\mu}, \widehat{\sigma})$ with $(\widehat{\mu}_0, \widehat{\sigma}_0)$, we obtain the FQs for μ and σ as

$$Q_\mu = \widehat{\mu}_0 - \frac{\widehat{\mu}^*}{\widehat{\sigma}^*}\widehat{\sigma}_0 \quad \text{and} \quad Q_\sigma = \frac{\widehat{\sigma}_0}{\widehat{\sigma}^*}. \tag{11}$$

For a given $(\widehat{\mu}_0, \widehat{\sigma}_0)$, the percentiles of Q_μ and Q_σ can be estimated by Monte Carlo simulation. Following the lines of the uncensored case in Section 2, we can obtain joint fiducial distribution empirically by generating samples from a $N(0,1)$ distribution with detection limits $DL_i^* = (\ln DL_i - \widehat{\mu}_0)/\widehat{\sigma}_0$, $i = 1, ..., k$. Calculation details for computing a UPL for the mean of a future sample is given in the following algorithm.

Algorithm 2

1. For a given sample with detection limits $DL_1, ..., DL_k$ from a lognormal distribution, compute the MLEs $\widehat{\mu}_0$ and $\widehat{\sigma}_0$ based on the log-transformed sample and log-transformed DLs.

2. Generate a sample of size n with detection limits $DL_i^* = (\ln DL_i - \widehat{\mu}_0)/\widehat{\sigma}_0$ from a $N(0,1)$ distribution

3. Compute the MLEs $\widehat{\mu}^*$ and $\widehat{\sigma}^*$ based on the sample generated in the preceding step.

4. Calculate $Q_\mu = \widehat{\mu}_0 - \frac{\widehat{\mu}^*}{\widehat{\sigma}^*}\widehat{\sigma}_0$ and $Q_\sigma = \widehat{\sigma}_0/\widehat{\sigma}^*$.

5. Generate a sample $Y_1^*, ..., Y_{n_*}^*$ f rom a normal distribution with mean Q_μ and standard deviation Q_σ, and compute $\bar{Y}_e^* = \frac{1}{n_*}\sum_{i=1}^{n_*}\exp(Y_i^*)$.

6. Repeat steps 2 – 5 for a large number of times, say, 10,000

7. The $100(1-\alpha)$ percentile of these 10,000 $\bar{Y}_e^*$'s is a $100(1-\alpha)\%$ UPL for the mean of a future sample of size n_*.

The above algorithm is coded in R and posted at www.ucs.louisiana.edu/∼kxk4695/. For a given sample with multiple detection limits, the posted R function computes the upper prediction limit for the mean of a future sample.

3.2. Coverage Studies

The coverage probabilities of 95% UPLs were estimated for the censored case along the lines for the uncensored case in Section 2.4. We carried out simulation studies for the samples that include a single detection limit or two detection limits. For the case of single detection limit, the coverage probabilities are reported in Table 3 for $(n, n_*) = (10, 5)$ and proportion of nondetects in the population is .7 or less. The coverage probabilities for the case of two detection limits are given in the same table when $(n, n_*) = (12, 4)$.

The estimated coverage probabilities, being very close to the nominal level .95, indicate that the fiducial UPLs are very satisfactory when the proportion of nondetects is .60 or less. For $(n, n_*) = (10, 5)$, the coverage probabilities are slightly less than the nominal level .95 when the proportion of nondetects $P_1 = .7$; however, for the same proportion of nondetects, these coverage probabilities are close to the nominal level when $(n, n_*) = (15, 5)$. For the case of two detection limits, the fiducial prediction limits are satisfactory when the overall proportion of nondetects is no more than .6. These simulation results indicate that the fiducial prediction limits could be liberal for small n and large proportion of nondetects. For a moderate sample size, say, 15 or more, the fiducial UPLs should work satisfactorily as long as the proportion of nondetects is no more than .70.

4. Examples

Example 1. To illustrate the application of the lognormal prediction limit for a future mean value, we shall use the example where the toxin of concern is lead (Bhaumik and Gibbons, 2004). An objective of this example is to determine whether a closed plating facility is safe for future industrial use. In the past, a portion of the facility might have been used as a waste disposal area. To determine the lead-impacted soil, $n_* = 5$ soil borings were installed. To check whether the on-site mean lead concentration at this area of the facility exceeds background, $n = 15$ off-site soil samples were collected in areas that were uninfluenced by the activities at the facility. The data are reported in Table 5 of Bhaumik and Gibbons (2004), and are reproduced here in Table 4. Bhaumik and Gibbons have also verified that the data satisfy lognormality assumption.

The mean and variance of the log-transformed background data are $\bar{x} = 2.1815$ and $s^2 = 2.3463$. Using their approach, Bhaumik and Gibbons (2004) have estimated the 95% upper prediction limit as 152.3 mg/kg. The 95% fiducial UPL using Algorithm 1 with 1,000,000 runs is obtained as 137.5 mg/kg. Martin and Lingham (2016) have obtained the UPL as 136.16 mg/kg on the basis of $\overline{G}_Y$ in (7). The little difference between their UPL and our UPL could be due to simulation error.

As the arithmetic mean of the on-site data 83.6 is less than the UPL, we conclude that the on-site concentrations do not significantly exceed those of the off-site. Finally, we note that the UPL based on the Bhaumik and Gibbons' approach is considerably greater than the fiducial UPL, and this comparison is consistent with our simulation studies in Section 3.2 where we observed that Bhaumik and Gibbons' approach could be conservative for some sample size and parameter configurations.

Example 2. To find fiducial prediction limits for samples with multiple DLs, we consider the atrazine concentrations data from a series of Nebraska wells reported in Table 9.7 of Helsel (2005, p. 159). The original data were altered by adding a second detection limit at .05 (see Helsel, 2005, p. 229), and they are reproduced here in Table 5. The probability plot in Figure 5.5 of Helsel (2005) indicates that lognormality assumption is tenable.

For this data set, $n = 24$, $DL_1 = .01$, $DL_2 = .05$, the number of nondetects below DL_1

Table 3: Coverage probabilities of 95% UPL for the mean of a future sample of size n_* based on a background sample of size n

single detection limit

$n = 10, n_* = 5$

	(μ, σ)				(μ, σ)			
P_1	(1, .1)	(1, .5)	(1, 1)	(1, 2)	(2, .1)	(2, .5)	(2, 1)	(2, 2)
0.1	.948	.950	.956	.948	.950	.951	.950	.950
0.2	.950	.951	.951	.944	.951	.951	.948	.951
0.3	.950	.950	.947	.957	.950	.955	.945	.955
0.5	.940	.946	.944	.947	.945	.945	.951	.945
0.6	.939	.938	.941	.938	.933	.938	.940	.937
0.7	.925	.928	.926	.932	.927	.929	.933	.927
				$n = 15, n^* = 5$				
0.1	.944	.945	.949	.947	.951	.948	.949	.953
0.2	.948	.949	.953	.955	.948	.946	.950	.951
0.3	.952	.953	.950	.951	.942	.951	.944	.947
0.5	.945	.954	.954	.941	.952	.945	.940	.948
0.6	.934	.952	.942	.946	.949	.949	.951	.944
0.7	.944	.945	.946	.945	.944	.944	.951	.947

two detection limits

$n = 12, n_* = 4$

	(μ, σ)				(μ, σ)			
(P_1, P_2)	(1, .1)	(1, .5)	(1, 1)	(1, 2)	(2, .1)	(2, .5)	(2, 1)	(2, 2)
(.1, .2)	.954	.951	.956	.948	.947	.953	.949	.950
(.1, .3)	.952	.951	.951	.944	.955	.951	.951	.951
(.2, .2)	.945	.947	.947	.957	.954	.950	.954	.955
(.2, .3)	.951	.952	.944	.947	.945	.948	.949	.945
(.1, .5)	.947	.950	.951	.948	.947	.947	.955	.947
				$n = 16, n^* = 7$				
(.1, .2)	.950	.947	.953	.954	.947	.953	.949	.951
(.1, .3)	.951	.946	.950	.949	.955	.951	.947	.951
(.2, .2)	.947	.954	.954	.950	.954	.950	.954	.953
(.2, .3)	.948	.951	.947	.956	.945	.948	.951	.947
(.1, .5)	.951	.953	.946	.949	.947	.947	.950	.948

NOTE: P_i denotes the proportion of data below the detection limit DL_i

Table 4: Lead level (mg/kg) in soil boring samples in off-site and on-site locations

Off-site	26	63	3	70	16	5	1	57	5	3	24	2	1	48	3
On-site	50	82	95	103	88										

is $m_1 = 9$ and the number of nondetects below DL_2 is $m_2 = 2$. The MLEs based on the log-transformed data are $\widehat{\mu} = -4.206$ and $\widehat{\sigma} = 1.462$. Suppose it is desired to find a 95% UPL for the mean of a future sample of size 5. Using Algorithm 2 with 100,000 runs, we estimated the 95% UPL for a future sample of size 5 as 0.20. That is, the mean atrazine concentration of a future sample of size 5 is no more than .20 with confidence 95%.

Table 5: Atrazine concentrations (μg/L) in a series of Nebraska wells before June

.38	< .05	< .01	.03	.03	.05	.02	< .01	< .01	< .01	.11	.09
< .01	< .01	< .01	< .01	.02	< .05	.02	.02	.05	.03	.05	< .01

5. Concluding Remarks

We have used the fiducial approach to solve yet another important problem of predicting the mean of a future sample from a lognormal distribution. The fiducial method is not only simple, but also provides very accurate prediction limits. We also noted that the fiducial prediction limit is exact for predicting a single future observation. Our extensive simulation studies for very small sample sizes indicate that the fiducial UPLs are practically exact. However, proving the exactness of the approach theoretically seems to be difficult. An appealing feature of the fiducial approach is that it can be easily applied to find prediction limits for the mean of a future sample based on a sample with multiple detection limits. However, it should be noted that the fiducial approach to the censored case involves repeated calculation of the MLEs based on censored samples, which requires an efficient computational algorithm. In order to help statistical practitioners and other users, we have posted R codes at *www.ucs.louisiana.edu/∼kxk4695* to find UPLs based on a censored or uncensored samples.

Acknowledgements

The authors are grateful to an associate editor and two reviewers for providing useful comments and suggestions which enhanced the first version of the paper.

References

Bernarie MM (1971). "The Validity of the Log-normal Distribution of Pollutant Concentrations, Paper SU-18D, Proceedings of the 2nd International Clean Air Conference (CE-Trans-7589).

Bhaumik DK, Gibbons RD (2004). "An Upper Prediction Limit for the Arithmetic Mean of a Lognormal Random Variable." *Technometrics*, **46**, 239–248.

Dawid AP, Stone M (1982). "The functional-model basis of fiducial inference." *Annals of Statistics*, **10**, 1054–1074.

Efron B (1998). R.A. Fisher in the 21st century. *Statistical Science*, **13**, 95–122.

Esmen N, Hammad Y (1977). "Lognormality of Environmental Sampling Data." *Environmental Sci. Health* **A12**, 29–41.

Fisher RA. (1930) "Inverse Probability." *Proceedings of the Cambridge Philosophical Society*, **xxvi**, 528–535.

Fisher RA. (1935) "The Fiducial Argument in Statistical Inference." *Annals of Eugenics VI*, 91–98.

Georgopoulos PG, Seinfeld JH (1982). "Statistical distributions of air quality concentrations." *Environmental Science and Technology* 16:401–416.

Gilbert RO (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold Co., New York.

Hannig T, Iyer HK, Patterson P (2006). "Fiducial Generalized Confidence Intervals." *Journal of American Statistical Association*, **101**, 254 – 269.

Helsel DR. (2005). *Nondetects and Data Analysis*. Hoboken, NJ, Wiley.

Krishnamoorthy K, Mathew T (2003). "Inferences on the Means of Lognormal Distributions using Generalized P-values and Generalized ConïňĄdence Intervals." *Journal of Statistical Planning and Inference*, **115**, 103 – 121.

Krishnamoorthy K, Xu Z (2011). "Confidence Limits for Lognormal Percentiles and for Lognormal Mean based on Samples with Multiple Detection Limits." *Annals of Occupational Hygiene*, **55**, 495–509.

Krishnamoorthy K, Mathew T, Xu Z (2014). "Standardized Likelihood Inference for the Mean and Percentiles of a Lognormal Distribution based on Samples with Multiple Detection Limits." *Journal of Environmental Statistics*, **6**, 1–18.

Krishnamoorthy K, Wang X (2016). "Fiducial Inference on Gamma Distribution: Uncensored and Censored Cases. *Environmetrics*, **27**, 479 – 493.

Land CE (1971). "Confidence Intervals for the Linear Function of the Normal Mean and Variance." *Annals of the Mathematical Statistics* 42:1187–1205.

Lyles RH, Kupper LL (1996). On Strategies for Comparing Occupational Exposure Data to Limits. *American Industrial Hygiene Association Journal*, **57**, 6–15.

Martin R, Lingam, RT (2016). Prior-Free Probabilistic Prediction of Future Observations. *Technometrics*, **58**, 225–235.

Oldham P (1953). "The Nature of the Variability of Dust Concentrations at the Coal Face. *British Journal Industrial Medicine*, **10**, 227–234.

Ogden TL (2010). "Handling Results Below the Level of Detection." *Annals of Occupational Hygiene*, **54**, 255–256.

Selvin S, Rappaport SM (1989). "Note on the Estimation of the Mean Value from a Lognormal Distribution." *American Industrial Hygiene Association Journal*, **50**, 627–630.

Speer DR, Waite DA (1975). "Statistical Distributions as Applied to Environmental Surveillance Data." Battelle, Pacific Northwest Laboratories, BNWL–SA-5482.

Tsui KW, Weerahandi S (1989). "Generalized P-values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters." *Journal of American Statistical Association*, **84**, 602–607.

Wang CM, Hannig J, Iyer HK (2012). "Fiducial prediction intervals." *Journal of Statistical Planning and Inference*, **142**, 1980–1990.

Weerahandi S (1993). "Generalized Confidence Intervals." *Journal of American Statistical Association*, **88**, 899–905.

Zou GY, Taleban J, Huoc CY (2009). "Confidence Interval Estimation for Lognormal Data with Application to Health Economics." *Computational Statistics and Data Analysis*, **53**, 3755–3764.

Affiliation:

K. Krishnamoorthy
Department of Mathematics
University of Louisiana at Lafayette
Lafayette, LA 70508-1010
E-mail: krishna@louisiana.edu
URL: http://www.ucs.louisiana.edu/~kxk4695

Journal of Environmental Statistics http://www.jenvstat.org
Volume 8, Issue 7 Submitted: 2017-12-18
September 2018 Accepted: 2018-09-16