International Journal of Statistics and Probability; Vol. 4, No. 2; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Comparison of Two Means of Two Log-Normal Distributions When Data is Singly Censored

Abou El-Makarim A. Aboueissa[1]

[1] Department of Mathematics and Statistics, University of Southern Maine, USA

Correspondence: Abou El-Makarim A. Aboueissa, Department of Mathematics and Statistics, University of Southern Maine, 96 Falmouth Street, Portland, Maine 04104-9300, USA. E-mail: aaboueissa@usm.maine.edu

Received: March 3, 2015 Accepted: March 24, 2015 Online Published: April 7, 2015

doi:10.5539/ijsp.v4n2p73 URL: http://dx.doi.org/10.5539/ijsp.v4n2p73

Abstract

It is common in environmental and biomedical data analysis to deal with censored data that are log-normally distributed. This paper is concerned with the statistical analysis for comparing the means of two independent log-normal distributions from censored data with a single detection limit. The method of maximum likelihood will be used to obtain closed form estimates for population parameters under different hypotheses. A test procedure for comparing the means of two independent log-normal populations in the presence of censored data is also introduced and evaluated. Asymptotic chi-square test is used in the proposed test procedure. Worked example is given illustrating the use of the methods provided utilizing a computer program written in the R language. A simulation study was performed to examine the power of the proposed test procedure introduced in this article.

Keywords: detection limits, censored data, normal and log-normal distributions, maximum likelihood estimators, likelihood ratio test

1. Introduction

The processing of the analytical results of environmental data containing potentially hazardous chemicals is often complicated by the fact that some of these pollutants are present at trace levels which cannot be measured reliably and consequently are reported as results lying numerically below a detection limit, DL. In general, censoring means that observations at one or both tails are not available. Left-censored data commonly arise in environmental contexts. Left-censored data (data reported as less than detection limit) can occur when the substance or attribute being measured is either absent or exists at such low concentrations that the substance is not present above the DL level. Data sets containing left-censored observations are referred to as left-censored data. When more than two distinct detection limits DL_1, DL_2 , ..., DL_k ($k \geq 3$) are reported, the data are said to be multiply-left-censored, (USPEA 1989b). It is common to have environmental data contains detection limits. Left censoring frequently arises in environmental studies due to: (1) sometimes nondetect is reported because the measurement lies below a threshold set by the client or laboratory, (2) sometimes the instrumentation registers a low signal, but the chemist decides that "unpollutant" environmental samples could give a similar signal and reports nondetect instead of the observed measurement, (3) sometimes the signal produced by the pollutant is too small for the instrumentation to discriminate from background noise, or (4) sometimes a signal is registered, but certain criteria that identify the compound are not met. A sample for which some observations are known only to fall above a known detection limit, while the remaining observations falling below the detection limit are fully measured and reported is called right censored. This type of data are so common in biomedical studies. In many environmental applications the distribution of variables such as chemical concentration, inhalation, digestion, and consumption rates are positive and skewed to the right. Hence, censored observations occur between zero and DL. In some instances a log transformation can provide a more natural scale to analyze such measurements. Samples to be considered in this paper are those that are Type I single-left-censored. Suppose that a sample of n data points is given of which m data points are non-censored (fully measured), and the remaining $m_c = n - m$ observations are left-censored with a single detection limit DL. In such Type I censored samples DL is fixed, whereas m and m_c are random.

Nondetect values can cause an especially difficult problem when the goal is to compare two different populations. There has been a great deal of literature on the subject of the statistical inference of the parameters of normal and

log-normal populations from both fully measured and censored data. Gupta and Li (2006) developed a score test for testing the equality of the means of two independent log-normal populations from fully measured data. Zhou et al (1997) considered two methods for comparing the means of two independent log-normal non-censored samples. Harris (1991) considered two parametric and two non-parametric methods for testing the equality of medians of two independent log-normal distributions when some data are left-censored. Prentice (1978) developed linear rank tests with right censored data. Millard and Deverel (1988) adapted several existing right censored non-parametric procedure so that they can be used in environmental setting with left-censored data. Methods for the estimation of the log-normal parameters for one-sample cases where there may exist left-censored data are discussed by El-Shaarawi (1989). Stoline (1993) extended results first suggested by Harris (1991) and proposed a procedure for comparing medians of two independent log-normal distributions where some data may be left-censored. Stoline (1993)used the Expectation Maximization (EM) algorithm introduced by Dempster et al. (1977) to calculate the maximum likelihood estimates of population parameters μ and σ. Other suggested methods for estimating population parameters from censored samples are discussed in Marco (2005), Jin et al (2010), Gibbons (1994), Gleit (1985), El-Shaarawi and Esterby (1992), Elshaarawi and Dolan (1989), Gilbert (1987) and Schneider (1986).

The purpose of this paper is to provide a parametric procedure for comparing means of two independent log-normally distributed populations utilizing left-censored data sets. The method of maximum likelihood will be used to obtain estimates of population parameters μ and σ. To facilitate the application of this procedure, a computer program is written in the R language which calculates the maximum likelihood estimates, asymptotic chi-square test statistics and their p-values. A numerical example is given illustrating the use of this procedure utilizing a computer program written in the R language.

2. Assumptions and Notations

Assume that there exists two random samples of n_1 and n_2 data values: $y_{11}, y_{12}, ..., y_{1m_1}, y_{1m_1+1}, ..., y_{1n_1}$ and $y_{21}, y_{22}, ..., y_{2m_2}, y_{2m_2+1}, ..., y_{2n_2}$ taken from two independent log-normal populations $LN(\mu_1, \sigma_1)$ and $LN(\mu_2, \sigma_2)$, respectively. Where $LN(\mu, \sigma)$ denotes a log-normally distributed variable y with the probability density function

$$f(y; \mu, \sigma) = \frac{1}{y\,\sigma\sqrt{2\pi}}\, e^{-\frac{(\log y-\mu)^2}{2\sigma^2}}, \quad \text{for } y > 0,$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. For convenience, for each sample i let us assume that the first m_i observations $y_{i1}, y_{i2}, ..., y_{im_i}$ are non-censored (fully measured) and the remaining $m_{c_i} = n_i - m_i$ observations are left-censored for $i = 1, 2$. It is assumed that each sample i has a detection limit LDL_i, for $i = 1, 2$, where LDL_i is the detection limit in lognormal sample i for $i = 1, 2$. For left censored observations, it is assumed that for each sample i it is only known that there are m_{c_i} observations less than LDL_i, for $i = 1, 2$. That is for sample i, $y_{ij} < LDL_i$ for $i = 1, 2$ and $j = 1, 2, ..., m_{c_i}$. The parameters for the i^{th} log-normal population can be expressed as functions of the parameters μ and σ as:

$$\begin{aligned} \textit{mean}: &\quad \mu_{y_i} = e^{\mu_i + \frac{\sigma_i^2}{2}} \\ \textit{medain}: &\quad M_{y_i} = e^{\mu_i} \\ \textit{variance}: &\quad \sigma^2_{y_i} = \gamma_i(\gamma_i - 1)\, e^{2\mu_i} \\ \textit{skewness}: &\quad s_{y_i} = (\gamma_i + 2)\sqrt{(\gamma_i - 1)} \end{aligned}$$

where $\gamma_i = e^{\sigma_i^2}$ for $i = 1, 2$.

Let

$$x_{ij} = \begin{cases} \ln(y_{ij}), & \text{for } i = 1, 2 \text{ and } j = 1, 2, ..., m_i, \\ DL_i = \ln(LDL_i), & \text{for } i = 1, 2 \text{ and } j = 1, 2, ..., m_{c_i}. \end{cases}$$

where LDL_i is the detection limit in the i^{th} lognormal sample.

To simplify the presentation in this paper, the analysis is described and illustrated by reference to the analysis of normally distributed data, though this condition occurs infrequently in typical environmental data analysis. However, it is frequently necessary to transform real environmental data before analysis; typically the logarithmic transformation of $x_{ij} = \ln(y_{ij})$ is used, although other transformations are possible. When the logarithmic or other transformation is used prior to censored data set analysis, it is necessary to transform the analysis results back to

the original scale of measurement following parameter estimation. For each sample i let

$$\bar{x}_{m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}\ , \quad \text{and} \quad s^2_{m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_{m_i})^2$$

be the sample mean and sample variance of the m_i non-censored observations $x_{i1}, x_{i2}, ..., x_{im_i}$, for $i = 1, 2$. Let the functions $\phi(.)$ and $\Phi(.)$ be the pdf and cdf of the standard unit normal. Define

$$\Phi(\xi_i) = \int_{-\infty}^{\xi_i}\phi(t)dt\ , \quad \text{where}\ \ \xi_i = \frac{DL_i - \mu_i}{\sigma_i}\ \text{for}\ i = 1, 2,$$

and

$$\Phi(\xi) = \int_{-\infty}^{\xi}\phi(t)dt\ , \quad \text{where}\ \ \xi = \frac{DL - \mu}{\sigma}\ .$$

We also define

$$\overline{DL} = \frac{DL_1 + DL_2}{2}\ \ ,\ \ \bar{\xi} = \frac{\xi_1 + \xi_2}{2}\ \ \text{and}\ \ W(x) = \frac{\phi(x)}{\Phi(x)}.$$

The likelihood function of the samples under consideration is given by:

$$L(\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^{2}\left(\frac{n_i!}{m_i!\ m_{c_1}!}[\Phi(\xi_i)]^{m_{c_i}}[\frac{1}{\sigma_i\sqrt{(2\pi)}}]^{m_i}\ e^{-\frac{1}{2\sigma_i^2}\sum_{j=1}^{m_i}(x_{ij}-\mu_i)^2}\right) \tag{2.1}$$

The two log-normal population means are confirmed equal whenever the null hypothesis $H_{0LN} : \mu_{y_1} = \mu_{y_2}$ is accepted in favor of the alternative hypothesis $H_{ALN} : \mu_{y_1} \neq \mu_{y_2}$ or equivalently whenever the null hypothesis $H_{0N} : \mu_1 = \mu_2 = \mu$ *and* $\sigma_1 = \sigma_2 = \sigma$ (overall homogeneity) is accepted in favor of one of the alternative hypotheses: $H_{A_1N} : \mu_1 \neq \mu_2$ *and* $\sigma_1 \neq \sigma_2$ (overall heterogeneity), $H_{A_2N} : \mu_1 \neq \mu_2$ *and* $\sigma_1 = \sigma_2 = \sigma$ (mean heterogeneity, variance homogeneity), or $H_{A_3N} : \mu_1 = \mu_2 = \mu$ *and* $\sigma_1 \neq \sigma_2$ (mean homogeneity, variance heterogeneity). For the sake of simplicity the test procedure for the null hypothesis H_{0N} versus the alternative hypothesis H_{A_1N} will be considered in this paper. The method of maximum likelihood will be used to obtain the maximum likelihood estimates of population parameters under the hypotheses H_{0N} and H_{A_1N}.

3. Maximum Likelihood Estimates of Population Parameters

In this section the maximum likelihood estimates of population parameters μ_i and σ_i, for $i = 1$ *and* 2, are derived under each of the hypotheses H_{0N} and H_{A_1N}. The derivation of these estimates is now described.

3.1 Maximum Likelihood Estimates under H_0N

Under the hypothesis H_{0N}, x_{ij}, for $i = 1, 2$ and $j = 1, 2, ..., n_i$, are assumed to be normally distributed with mean μ and standard deviation σ. That is, it is assumed that there exists a random sample of $n = n_1 + n_2$ data values taken from a normal population with mean μ and standard deviation σ. For convenience, let us assume that the first $m = m_1 + m_2$ observations are non-censored (fully measured) and the remaining $m_c = m_{c_1} + m_{c_2} = n - m$ observations are left-censored. For left censored observations, it is only known that m_{c_1} observations are reported as less than DL_1 and m_{c_2} observations are reported as less than DL_2.

Case 1: In this case it is assumed that $DL_1 \neq DL_2$. The likelihood function $L_{H_{0N}}(\mu, \sigma)$ under H_{0N} determined by the pooled sample $x_{11}, x_{12}, ..., x_{1m_1}, x_{2(m_1+1)}, ..., x_{2m}, x_{1(m+1)}, ..., x_{1(m+m_{c_1})}, x_{2(m+m_{c_1}+1)}, ..., x_{2n}$, where $m = m_1 + m_2$, $m_c = m_{c_1} + m_{c_2}$ and $n = n_1 + n_2 = m + m_c$, is given by:

$$L_{H_{0N}}(\mu, \sigma) = \prod_{i=1}^{2}\left(\frac{n_i!}{m_{c_i}!\ m_i!}\ [\Phi(\xi_i)]^{m_{c_i}}\ [\frac{1}{\sigma\sqrt{2\pi}}]^{m_i}\ e^{-\frac{1}{2\sigma^2}\ \sum_{j=1}^{m_i}(x_{ij}-\mu)^2}\right) \tag{3.1}$$

Hence, the corresponding log-likelihood function of (3.1) is given by:

$$\begin{aligned}\ell_{H_{0N}}(\mu, \sigma) = \ln\left(\prod_{i=1}^{2}\frac{n_i!}{m_{c_i}!\ m_i!}[2\pi]^{-\frac{m_i}{2}}\right) - \sum_{i=1}^{2} m_i\ \ln(\sigma) + \sum_{i=1}^{2} m_{c_i}\ \ln[\Phi(\xi_i)] \\ - \sum_{i=1}^{2}\frac{1}{2\ \sigma^2}\sum_{j=1}^{m_i}(x_{ij} - \mu)^2\end{aligned} \tag{3.2}$$

For convenience, define $h = \frac{m_c}{n}$, $h_i = \frac{m_{c_i}}{n}$, and $\frac{m_{c_i}}{m} = \frac{h_i}{1-h}$, for $i = 1, 2$. For the pooled sample let

$$\bar{x}_m = \frac{1}{m}\sum_{i=1}^{2}\sum_{j=1}^{m} x_{ij}\ , \quad and \quad s_m^2 = \frac{1}{m}\sum_{i=1}^{2}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_m)^2$$

be the sample mean and sample variance of the m non-censored observations $x_{11}, x_{12}, ...,$ $x_{1m_1}, x_{2(m_1+1)}, ..., x_{2m}$, respectively.

The maximum likelihood estimates for $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ are the solutions to equations (3.3) and (3.4), the partial derivatives for the log-likelihood equation with respect to μ and σ:

$$\frac{\partial \ell_{H_{0N}}(\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{2}\sum_{j=1}^{m_i}(\frac{x_{ij}-\mu}{\sigma}) - \sum_{i=1}^{2} m_{c_i}\frac{\phi(\xi_i)}{\Phi(\xi_i)} = 0 \tag{3.3}$$

$$\frac{\partial \ell_{H_{0N}}(\mu,\sigma)}{\partial \sigma} = \sum_{i=1}^{2}\sum_{j=1}^{m_i}(\frac{x_{ij}-\mu}{\sigma})^2 - \sum_{i=1}^{2} m_i - \sum_{i=1}^{2} m_{c_i}\xi_i\frac{\phi(\xi_i)}{\Phi(\xi_i)} = 0 \tag{3.4}$$

The expectation maximization (EM) algorithm will be used iteratively to obtain the solutions $\hat{\mu}$ and $\hat{\sigma}$ to the maximum likelihood equations (3.3) and (3.4). The EM algorithm was proposed by Dempster et. al. (1977) for calculating the maximum likelihood estimated from censored samples. The procedure consists of alternately estimating the censored observations from the current parameter estimates and estimating the parameters from the actual and estimated observations. The EM algorithm can be used to calculate the maximum likelihood estimates for the mean μ and standard deviation σ of a normal distribution from both singly- and multiply-censored samples. A brief description for the EM algorithm is given here.

At step 0 of the EM algorithm all non-censored observations are used to calculate the initial estimates of μ and σ as follows:

$$\hat{\mu}_0 = \bar{x}_m = \frac{1}{m}\sum_{i=1}^{2}\sum_{j=1}^{m} x_{ij}\ , \quad and \quad \hat{\sigma}_0^2 = s_m^2 = \frac{1}{m}\sum_{i=1}^{2}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_m)^2$$

Let $\hat{\mu}_s$ and $\hat{\sigma}_s$ be the maximum likelihood estimates of μ and σ at step s of this procedure. At step $s + 1$, each censored observation x_{ij} (where i=1,2; j=1,2,..., m_{c_i}) is replaced by an estimate of $\hat{\mu}_s - \hat{\sigma}_s W(\frac{x_{ij}-\hat{\mu}_s}{\hat{\sigma}_s})$.

Let the values u_{ij} be calculated at step $s + 1$ as follows:

$$u_{ij} = \begin{cases} x_{ij}\ , & \text{for non-censored data values} \\ \hat{\mu}_s - \hat{\sigma}_s W(\frac{x_{ij}-\hat{\mu}_s}{\hat{\sigma}_s})\ , & \text{for censored data values} \end{cases}$$

So the updated estimates $\hat{\mu}_{s+1}$ and $\hat{\sigma}_{s+1}$ of μ and σ are given by

$$\hat{\mu}_{s+1} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{m_i} u_{ij} + \sum_{i=1}^{2}\sum_{j=1}^{m_{c_i}} u_{ij}}{n}$$

and

$$\hat{\sigma}_{s+1}^2 = \frac{\sum_{i=1}^{2}\sum_{j=1}^{m_i}(u_{ij} - \hat{\mu}_{s+1})^2 + \sum_{i=1}^{2}\sum_{j=1}^{m_{c_i}}(u_{ij} - \hat{\mu}_{s+1})^2}{\sum_{i=1}^{2} m_i + \sum_{i=1}^{2}\sum_{j=1}^{m_{c_i}}\gamma(\frac{x_{ij}-\hat{\mu}_s}{\hat{\sigma}_s})}$$

where the function $\gamma(t)$ is defined as:

$$\gamma(t) = W(t)(W(t) + t) \quad \text{and} \quad W(t) = \frac{\phi(t)}{\Phi(t)}$$

More details about the EM algorithm procedure can be found in Wolynetz (1979). Convergence is achieved if both

$|\hat{\mu}_s - \hat{\mu}_{s+1}| < 0.00001$ and $|\hat{\sigma}_s - \hat{\sigma}_{s+1}| < 0.00001$ occur. When these convergence criteria are met, the maximum likelihood estimates for μ and σ are then given by $\hat{\mu} = \hat{\mu}_s$ and $\hat{\sigma} = \hat{\sigma}_s$, respectively.

Case 2: In this case it is assumed that $DL_1 = DL_2 = DL$. The likelihood function $L_{H_{0N}}(\mu, \sigma)$ under H_{0N} determined by the pooled sample of $n = n_1 + n_2$ observations of which $m_c = m_{c_1} + m_{c_2}$ are left censored and share the same detection limit DL, is given by:

$$L_{H_{0N}}(\mu, \sigma) = \frac{n!}{m!\ m_c!}\ [\Phi(\xi)]^{m_c}\ \left[\frac{1}{\sigma\ \sqrt{2\pi}}\right]^{m}\ e^{\frac{1}{2\sigma^2}\sum_{j=1}^{m}(x_j-\mu)^2} \tag{3.5}$$

Hence, the corresponding log-likelihood function of (3.10) is given by:

$$\ell_{H_{0N}}(\mu, \sigma) = ln\left(\frac{n!}{m!\ m_c!}\ [2\pi]^{-\frac{m}{2}}\right) + m_c\ ln[\Phi(\xi)] - m\ ln(\sigma) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(x_j - \mu)^2 \tag{3.6}$$

The maximum likelihood estimates for $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ are the solutions to equations (3.5) and (3.6), the partial derivatives for the log-likelihood equation with respect to μ and σ:

$$\frac{\partial \ell_{H_{0N}}(\mu, \sigma)}{\partial \mu} = -m_c\ \frac{\phi(\xi)}{\Phi(\xi)} + \sum_{j=1}^{m}(\frac{x_j - \mu}{\sigma}) = 0 \tag{3.7}$$

$$\frac{\partial \ell_{H_{0N}}(\mu, \sigma)}{\partial \sigma} = \sum_{j=1}^{m}(\frac{x_j - \mu}{\sigma})^2 - m - m_c\ \xi\ \frac{\phi(\xi)}{\Phi(\xi)} = 0 \tag{3.8}$$

By solving equations (3.7) and (3.8) for μ and σ lead to the following maximum likelihood estimating equations:

$$\mu = \bar{x}_m - \lambda_0\ (\bar{x}_m - DL\ )\ , \tag{3.9}$$

$$\sigma = \sqrt{s_m^2 + \lambda_0\ (\bar{x}_m - DL\ )^2}\ , \tag{3.10}$$

and

$$\gamma = \frac{\left[1 - (\frac{h}{1-h})\ Z(\xi)\left((\frac{h}{1-h})\ Z(\xi) - \xi\right)\right]}{\left[(\frac{h}{1-h})\ Z(\xi) - \xi\right]^2}, \tag{3.11}$$

where

$$\lambda_0 = \frac{(\frac{h}{1-h})\ Z(\xi)}{(\frac{h}{1-h})\ Z(\xi) - \xi}\ , \tag{3.12}$$

and

$$\gamma = \frac{s_m^2}{[\bar{x}_m - DL\ ]^2}\ .$$

To obtain the desired maximum likelihood estimates of μ and σ from (3.9)-(3.12), it is necessary to estimate the auxiliary function $\lambda = \lambda(\xi, h)$. To obtain the estimate value $\hat{\lambda}$ of λ, it is necessary to solve the rather complex non-linear estimating equation (3.11) for the estimate $\hat{\xi}$ of ξ. Because of the difficulty of solving this equation explicitly for ξ, an iterative method proposed by Aboueissa and Stoline (2004) will be used for obtaining the maximum likelihood estimates $\hat{\mu}_0$ and $\hat{\sigma}_0$ of μ and σ. Let $\hat{\xi}$ be the estimate of ξ. Thus the maximum likelihood estimates $\hat{\mu}_0$ and $\hat{\sigma}_0$ of μ and σ are given by:

$$\hat{\mu}_0 = \bar{x}_m - \hat{\lambda}_0\ (\bar{x}_m - DL\ )\ , \tag{3.13}$$

and

$$\hat{\sigma}_0 = \sqrt{s_m^2 + \hat{\lambda}_0\ (\bar{x}_m - DL\ )^2}\ , \tag{3.14}$$

where

$$\hat{\lambda}_0 = \lambda_0(\hat{\xi}, h)\ .$$

Alternatively, the EM algorithm estimation procedure presented in case 1 can be used to obtain the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ for μ and σ using equations (3.7) and (3.8).

3.2 Maximum Likelihood Estimates under H_A1N

Under the hypothesis H_{A1N} x_{ij} are assumed to be normally distributed with mean μ_i and standard deviation σ_i, for $i = 1, 2$ and $j = 1, 2, ..., n_i$. Thus the likelihood function under H_{A1N} is given by:

$$L_{H_{A1N}}(\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^{2}\left(\frac{n_i!}{m_{c_i}!\, m_i!}\ [\Phi(\xi_i)]^{m_{c_i}}\ [\frac{1}{\sigma_i\sqrt{2\pi}}]^{m_i}\ e^{-\frac{1}{2\sigma_i^2}\ \sum_{j=1}^{m_i}(x_{ij}-\mu_i)^2}\right) \tag{3.15}$$

Hence, the corresponding log-likelihood function of (3.15) is given by:

$$\begin{aligned}\ell_{H_{A1N}}(\mu_1, \mu_2, \sigma_1, \sigma_2) = \ln\left(\prod_{i=1}^{2}\frac{n_i!}{m_{c_i}!\, m_i!}[2\pi]^{-\frac{m_i}{2}}\right) - \sum_{i=1}^{2} m_i\ \ln(\sigma_i) + \sum_{i=1}^{2} m_{c_i}\ \ln[\Phi(\xi_i)] \\ - \sum_{i=1}^{2}\frac{1}{2\ \sigma_i^2}\sum_{j=1}^{m_i}(x_{ij}-\mu_i)^2\end{aligned} \tag{3.16}$$

The maximum likelihood estimates for $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i are the solutions to equations (3.17) and (3.18) for $i = 1, 2$.

$$\frac{\partial \ell_{H_{1N}}(\mu_i, \sigma_i)}{\partial \mu_i} = -m_{c_i}\ \frac{\phi(\xi_i)}{\Phi(\xi_i)} + \sum_{j=1}^{m_i}(\frac{x_{ij}-\mu_i}{\sigma_i}) = 0 \tag{3.17}$$

$$\frac{\partial \ell_{H_{10N}}(\mu_i, \sigma_i)}{\partial \sigma_i} = \sum_{j=1}^{m_i}(\frac{x_{ij}-\mu_i}{\sigma_i})^2 - m_i - m_{c_i}\ \xi_i\ \frac{\phi(\xi_i)}{\Phi(\xi_i)} = 0 \tag{3.18}$$

By solving equations (3.17) and (3.18) for μ_i and σ_i lead to the following maximum likelihood estimating equations:

$$\mu_i = \bar{x}_{m_i} - \lambda_i\ (\bar{x}_{m_i} - DL_i\ )\ , \tag{3.19}$$

$$\sigma_i = \sqrt{s_{m_i}^2 + \lambda_i\ (\bar{x}_{m_i} - DL_i\ )^2}\ , \tag{3.20}$$

and

$$\gamma_i = \frac{\left[1 - (\frac{h_i}{1-h_i})\ Z(\xi_i)\left((\frac{h_i}{1-h_i})\ Z(\xi_i) - \xi_i\right)\right]}{\left[(\frac{h_i}{1-h_i})\ Z(\xi_i) - \xi_i\right]^2}, \tag{3.21}$$

where

$$\lambda_i = \frac{(\frac{h_i}{1-h_i})\ Z(\xi_i)}{(\frac{h_i}{1-h_i})\ Z(\xi_i) - \xi_i}\ , \tag{3.22}$$

and

$$\gamma_i = \frac{s_{m_i}^2}{[\bar{x}_{m_i} - DL_i\ ]^2}\ .$$

for $i = 1, 2$, where $h_i = \frac{m_{c_i}}{n_i}$.

To obtain the desired maximum likelihood estimates of μ_1 and σ_1 from (3.19)-(3.22), it is necessary to estimate the auxiliary functions $\lambda_i(\xi_i, h_i)$. To obtain the estimate value $\hat{\lambda}_i$ of λ_i, it is necessary to solve the rather complex non-linear estimating equation (3.21) for estimates $\hat{\xi}_i$ of ξ_i for $i = 1, 2$. Because of the difficulty of solving this equation explicitly for ξ_i, an iterative method proposed by Aboueissa and Stoline (2004) will be used for obtaining the maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i, for $i = 1, 2$. Let $\hat{\xi}_i$ be the estimates of ξ_i for $i = 1, 2$. Thus the maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i under the hypothesis H_{A1N} are given by:

$$\hat{\mu}_i = \bar{x}_{m_i} - \hat{\lambda}_i\ (\bar{x}_{m_i} - DL_i\ )\ , \tag{3.23}$$

and

$$\hat{\sigma}_i = \sqrt{s_{m_i}^2 + \hat{\lambda}_i\ (\bar{x}_{m_i} - DL_i\ )^2}\ , \tag{3.24}$$

where

$$\hat{\lambda}_i = \lambda_i(\hat{\xi}_i, h_i) \ .$$

for $i = 1, 2$.

Alternatively, the EM algorithm estimation procedure presented above can be used to obtain the maximum likelihood estimates $\hat{\mu}_1$ and $\hat{\sigma}_1$ for μ_1 and σ_1 using equations (3.17) and (3.18) for $i = 1$. Similarly, the maximum likelihood estimates $\hat{\mu}_2$ and $\hat{\sigma}_2$ for μ_2 and σ_2 using equations (3.17) and (3.18) for $i = 2$.

4. Asymptotic Chi-Square Test

The estimated log-likelihood functions $\hat{\ell}_{H_{0N}}$ and $\hat{\ell}_{H_{A1N}}$ under the hypotheses H_{0N} (Case 1 and Case 2)and H_{A_1N}, respectively; are obtained by replacing population parameters by their maximum likelihood estimates. Therefore from (3.2), (3.6) and (3.16) we get:

$\hat{\ell}_{H_{0N}}$ **C**ase 1: $DL_1 \neq DL_2$

$$\begin{aligned}\hat{\ell}_{H_{0N}} &= \ell_{H_{0N}}(\hat{\mu}, \hat{\sigma}) \\ &= \ln\left(\prod_{i=1}^{2} \frac{n_i!}{m_{c_i}!\ m_i!}[2\pi]^{-\frac{m_i}{2}}\right) - \sum_{i=1}^{2} m_i\ \ln(\hat{\sigma}) + \sum_{i=1}^{2} m_{c_i}\ \ln[\Phi(\hat{\xi}_i)] \\ &\quad - \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{m_i}(\frac{x_{ij} - \hat{\mu}}{\hat{\sigma}})^2 \ ,\end{aligned} \tag{4.1}$$

$\hat{\ell}_{H_{0N}}$ **C**ase 2: $DL_1 = DL_2 = DL$

$$\begin{aligned}\hat{\ell}_{H_{0N}} &= \ell_{H_{0N}}(\hat{\mu}, \hat{\sigma}) \\ &= \ln\left(\frac{n!}{m!\ m_c!}\ [2\pi]^{-\frac{m}{2}}\right) + m_c\ \ln[\Phi(\hat{\xi})] - m\ \ln(\hat{\sigma}) \\ &\quad - \frac{1}{2}\sum_{j=1}^{m}(\frac{x_j - \hat{\mu}}{\hat{\sigma}})^2 \ ,\end{aligned} \tag{4.2}$$

and

$$\begin{aligned}\hat{\ell}_{H_{A1N}} &= \ell_{H_{A1N}}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) \\ &= \ln\left(\prod_{i=1}^{2} \frac{n_i!}{m_{c_i}!\ m_i!}[2\pi]^{-\frac{m_i}{2}}\right) - \sum_{i=1}^{2} m_i\ \ln(\hat{\sigma}_i) + \sum_{i=1}^{2} m_{c_i}\ \ln[\Phi(\hat{\xi}_i)] \\ &\quad - \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{m_i}(\frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i})^2\end{aligned} \tag{4.3}$$

Asymptotic $\alpha-$level chi-square test will be used to test the equality of the means of two independent log-normal populations. Asymptotic $\alpha-$level chi-square test to test the null hypothesis $H_{0LN} : \mu_{y_1} = \mu_{y_2}$ versus the alternative $H_{ALN} : \mu_{y_1} \neq \mu_{y_2}$ or equivalently to test the null hypothesis $H_{0N} : \mu_1 = \mu_2 = \mu$ *and* $\sigma_1 = \sigma_2 = \sigma$ versus the alternative hypotheses $H_{A_1N} : \mu_1 \neq \mu_2$ *and* $\sigma_1 \neq \sigma_2$ is defined by:

$$\chi_0^2 = -2(\hat{\ell}_{H_{0N}} - \hat{\ell}_{H_{A1N}}) > \chi^2_{(\alpha,2)} \tag{4.4}$$

where $\chi^2_{(\alpha,2)}$ is the upper $\alpha-$point for a chi-square random variable with 2 degrees of freedom. The p-value of this test statistic is defined by:

$$p - value = P(\chi^2_{(2)} \ > \ \chi^2_0). \tag{4.5}$$

Thus the null hypothesis that the means of two independent log-normal populations are equal will be rejected if $\chi_0^2 > \chi^2_{(\alpha,2)}$ or equivalently if $p - value < \alpha$.

Computer Programs: To facilitate the application of parameter estimation method described in this article, a computer program called "*Abou.Two.Lognormal.Estimates*" is written in the R language to automate parameters estimation from left-censored data sets that are normally or log-normally distributed and to obtain the estimated values of the log-likelihood functions under the hypotheses H_{0N} and H_{A_1N}. In addition, this computer program will be used to obtain the asymptotic $\alpha-$level chi-square test statistic and its p-value. Copy of source code is given in the Appendix section.

5. Case Study Example

Millard and Deverel (1988) compared the median copper and zinc trace element concentrations in groundwater sampled from two geological areas in the San Joaquin Valley, the Basin-Trough Zone and the Alluvial Fan Zone in California. Data from both sites are given in Table 1.

Table 1. Millard and Deverel copper and zinc data: Groundwater concentrations of copper and zinc at two geological zones in the San Joaquin valley, California

copper	Alluvial Fan Zone $n = 65$	< 1 < 1 3 3 5 1 4 4 2 2 1 2 < 5 11 < 1 2 2 2 2 < 20 2 2 3 3 < 20 < 10 7 5 2 2 < 10 7 12 < 1 20 16 < 5 1 2 < 5 3 2 8 7 5 < 5 2 < 10 < 5 < 5 2 10 2 4 < 5 2 3 9 < 5 2 2 2 2 1 1
	Basin-Trough Zone $n = 49$	2 2 12 2 1 < 10 < 10 4 < 10 < 1 1 < 2 < 2 1 2 < 10 3 < 1 1 1 3 < 5 17 23 9 9 3 3 < 15 < 5 4 < 5 < 5 < 5 4 8 1 15 3 3 1 6 3 6 3 4 5 14 4
zinc	Alluvial Fan Zone $n = 67$	< 10 9 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
	Basin-Trough Zone $n = 50$	20 10 60 20 12 8 < 10 14 < 10 17 < 3 11 5 12 4 3 6 3 15 13 4 20 20 70 60 40 30 40 17 10 20 20 5 10 50 30 25 10 < 10 40 20 10 20 20 30 20 30 50 90 20

These data sets contains seven distinct detection limits (LDL : 1 , 2 , 3 , 5 , 10 , 15 , 20) with censoring level (percentage of non-detected observations) of 22%. Millard and Deverel (1988) give three possible causes for multiple left-censoring when measuring the concentration of copper and/or zinc in shallow groundwater. First cause may be decreasing detection limits over time as measurement devices improve. Second, there may be more than one method available, and each method may be optimal in different ranges of zinc and/or copper concentration. A third cause involves the amount of dilution that a lab technician may use. In this article it is assumed that both data sets are singly left-censored, thus to utilize the estimation methods described here, the left-censored observations within each data set will be set to equal to the average of detection limits. Table 2 contains estimates of the normal and log-normal population parameters. It is noted that the highest reported concentration of zinc (620) in the Alluvial Fan Zone seems to be unusual data value since the second highest observed zinc concentration is (50) in this zone. Two different estimates of the normal and log-normal population parameters for the zinc data sets are reported. The first estimate includes all data and the second includes all data with the zinc data value 620 removed ($ZincW620R$). The corresponding estimates of the zinc means $\mu_{y_i} = e^{\mu_i + \frac{\sigma_i^2}{2}}$ and medians $m_{y_i} = e^{\mu_i}$ are also included in Table 2. The influence of the single large zinc data value 620 can be most clearly seen by comparing the estimates for σ_1 under the hypothesis H_{A1N}. The estimate is $\hat{\sigma}_1 = 0.887$ with the 620 value included and $\hat{\sigma}_1 = 0.674$ with the 620 value removed. The estimates for μ_1 with the 620 value included and with the 620 value removed do not differ appreciably. These estimates for μ_1 are $\hat{\mu}_1 = 2.405$ and $\hat{\mu}_1 = 2.382$ for these two cases, respectively. The corresponding estimates of the log-normal median zinc concentration in the Alluvial Fan Zone are $\hat{m}_{y_1} = 11.078$ with the 620 value included and $\hat{m}_{y_1} = 10.827$ with the 620 value removed. The comparable estimates of the log-normal mean zinc concentration are $\hat{\mu}_{y_1} = 16.418$ with the 620 value included and $\hat{\mu}_{y_1} = 13.587$ with the 620 value removed. In addition, the estimates of the log-normal standard deviation for zinc concentration in the Alluvial Fan Zone are $\hat{\sigma}_{y_1} = 14.7580$ with the 620 value included and $\hat{\sigma}_{y_1} = 10.343$ with the 620 value removed. The estimates of the log-normal median are similar, but the estimates of the log-normal mean and standard deviation are appreciably different, owing to the influence of the single large zinc data value 620. Table 2 contains the p-value results associated with the application of the recommended asymptotic chi-square test to the the Millard and Deverel (1988) copper and zinc data presented in Table 1. The Millard and Deverel (1988) p-value results using the normal scores permutation variance ($NS2P$) procedure are also presented in Table 2.

Table 2. Estimates of normal and log-normal parameter values from the copper and zinc data given in Table 1

	Copper		Zinc		ZincW620R	
Hypothesis	Alluvial Fan Zone	Basin-Trough Zone	Alluvial Fan Zone	Basin-Trough Zone	Alluvial Fan Zone	Basin-Trough Zone
	Estimations of Normal Parameters					
H_{0N}	$\hat{\mu}_0 = 1.108$ $\hat{\sigma}_0 = 0.773$		$\hat{\mu}_0 = 2.524$ $\hat{\sigma}_0 = 0.910$		$\hat{\mu}_0 = 2.501$ $\hat{\sigma}_0 = 0.819$	
H_{A1N}	$\hat{\mu}_1 = 1.077$ $\hat{\sigma}_1 = 0.717$	$\hat{\mu}_2 = 1.144$ $\hat{\sigma}_2 = 0.842$	$\hat{\mu}_1 = 2.405$ $\hat{\sigma}_1 = 0.887$	$\hat{\mu}_2 = 2.683$ $\hat{\sigma}_2 = 0.918$	$\hat{\mu}_1 = 2.382$ $\hat{\sigma}_1 = 0.674$	$\hat{\mu}_2 = 2.683$ $\hat{\sigma}_2 = 0.918$
	Estimations of Log-normal Parameters					
H_{0LN}	$\hat{\mu}_{0y} = 4.0833$ $\hat{m}_{0y} = 3.029$ $\hat{\sigma}_{0y} = 3.339$		$\hat{\mu}_{0y} = 18.879$ $\hat{m}_{0y} = 12.478$ $\hat{\sigma}_{0y} = 17.324$		$\hat{\mu}_{0y} = 16.929$ $\hat{m}_{0y} = 12.195$ $\hat{\sigma}_{0y} = 14.255$	
H_{ALN}	$\hat{\mu}_{y_1} = 3.796$ $\hat{m}_{y_1} = 2.936$ $\hat{\sigma}_{y_1} = 2.976$	$\hat{\mu}_{y_2} = 4.479$ $\hat{m}_{y_2} = 3.139$ $\hat{\sigma}_{y_2} = 3.869$	$\hat{\mu}_{y_1} = 16.418$ $\hat{m}_{y_1} = 11.078$ $\hat{\sigma}_{y_1} = 14.758$	$\hat{\mu}_{y_2} = 22.295$ $\hat{m}_{y_2} = 14.629$ $\hat{\sigma}_{y_2} = 20.609$	$\hat{\mu}_{y_1} = 13.587$ $\hat{m}_{y_1} = 10.827$ $\hat{\sigma}_{y_1} = 10.343$	$\hat{\mu}_{y_2} = 22.295$ $\hat{m}_{y_2} = 14.629$ $\hat{\sigma}_{y_2} = 20.609$
Test	The Asymptotic Chi-square Test: χ_0^2 (P-value)					
H_{0N} vs. H_{A1N}	1.5016 (0.2204)		2.7324 (0.0983)		8.8602 (0.0029)	
Test	The P-value of the Median equality test, Millard and Deverel (1988)					
$m_{y_1} = m_{y_2}$	0.320		0.020		– – – – –	

ZincW620R: Alluvial Fan Zone zinc data set with the data value 620 removed.

Copper Case: The p-value of the asymptotic chi-square test statistic of testing the null hypothesis H_{0N} versus H_{A1N} or equivalently H_{0LN} versus H_{ALN} is 0.2204. Therefore the hypothesis of equal means is accepted for copper at significance level of $\alpha = 0.05$. The reported p-value for equality of medians of Millard and Deverel $NS\,2P$ is 0.320.

Zinc Case: The p-value of the asymptotic chi-square test statistic of testing the null hypothesis H_{0N} versus H_{A1N} or equivalently H_{0LN} versus H_{ALN} is 0.0983 with the 620 value included. Therefore with the 620 value included the hypothesis of equal means is accepted for zinc at significance level of $\alpha = 0.05$. The p-vale of testing the null hypothesis H_{0N} versus H_{A1N} or equivalently H_{0LN} versus H_{ALN} is 0.0029 with the 620 value removed. Therefore with the 620 value removed the hypothesis of equal means is rejected for zinc at significance level of $\alpha = 0.05$. The reported p-value for equality of medians of Millard and Deverel $NS\,2P$ is 0.020.

6. Simulation Study

In this simulation study, type I error rates and power of the test procedure introduced in this article are investigated. A computer program was written in the R language for this purpose. For each combination of the population parameters μ_1 , μ_2, σ_1 and σ_2 described below, two sample size cases were considered: in case one, $n_1 = n_1 = 25$ and in the second case, $n_1 = n_2 = 75$. The first case will be referred to as the small sample size case and the second as the large sample size case.Censoring at two different detection limits was used for each case. The simulation study was performed with 10,000 repetitions ($N = 10,000$) of sample normal distributions for each combinations of $n, \mu_1, \mu_2, \sigma_1, \sigma_2$, and censoring level(s). Censoring levels were set at the 15th and 30th percentiles of the parent distribution(s). In order to check the Type I error, the population parameters were specified as $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 1$ as shown in Table 3. In order to check the power, the population parameters were specified as $\mu_1 = 0$, $\mu_2 = 0.1(0.1)1.0$, $\sigma_1 = 1$, and $\sigma_2 = 1.0(0.1)2.0$ as shown in Table 4.

Table 3. The Estimated Simulated Type I Error Rates: $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1$

Sample Size	Censoring Level	Estimated α
Small ($n = 25$)	15%	0.067
Small ($n = 25$)	30%	0.063
Large ($n = 75$)	15%	0.054
Large ($n = 75$)	30%	0.056

The following observations and conclusions are made from an examination of the simulation results reported in Tables 3 and 4.

From Table 3, one can see that the estimated simulated Type I error rates are slightly higher than 0.06 (0.067, 0.063) for the small sample size case, and slightly higher than 0.05 (0.054, 0.056) for the large sample size case. The censoring levels do not seem to affect the value of Type I error rate, α.

Table 4. The Estimated Simulated Power Rates

	Small Sample Size (n = 25)		Large Sample Size (n = 75)	
$(\mu_1, \mu_2, \sigma_1, \sigma_2)$	Censoring Level (15%)	Censoring Level (30%)	Censoring Level (15%)	Censoring Level (30%)
(0, 0.1, 1, 1.1)	0.1893	0.1763	0.2714	0.2657
(0, 0.2, 1, 1.2)	0.3187	0.2748	0.5348	0.5749
(0, 0.3, 1, 1.3)	0.4615	0.4158	0.8357	0.8143
(0, 0.4, 1, 1.4)	0.5834	0.5857	0.9518	0.9435
(0, 0.5, 1, 1.5)	0.7443	0.7246	0.9947	0.9858
(0, 0.6, 1, 1.6)	0.8387	0.8195	0.9994	0.9986
(0, 0.7, 1, 1.7)	0.9035	0.8963	1.0000	1.0000
(0, 0.8, 1, 1.8)	0.9526	0.9375	1.0000	1.0000
(0, 0.9, 1, 1.9)	0.9628	0.9624	1.0000	1.0000
(0, 1.0, 1, 2.0)	0.9893	0.9758	1.0000	1.0000

From Table 4, one can see that the estimated simulated power is higher for large sample size case than the small sample size case, and slightly higher for the lower level of censoring. Specifically, in the small sample size case with 15% (30%) censoring level we reach a power of 0.9893 (0.9758) when the difference between μ_1 and μ_2 is 1.0, and the difference between σ_1 and σ_2 is 1.0. Alternatively, in the large sample size case with 15% (30%) censoring we reach a power above 0.99 when the difference between μ_1 and μ_2 is 0.6, and the difference between σ_1 and σ_2 is 0.6; and a power of 1.0 when the difference between $\mu's$ and $\sigma's$ is 0.7.

In summary, the test procedure introduced in this article maintains its stated significance level and has much power with larger sample size and a bit less power with greater censoring levels. In addition, the power decreases when the censoring level moves from 0.15 to 0.30. Also, the power increases greatly when the sample size moves from the order of 25 to the order of 75.

7. Conclusions and Remarks

It is well known that the log-normal distribution is widely used in modeling environmental and biomedical censored data. This article has dealt with the problem of comparing means of two independent log-normal populations in the presence of singly left-censored data. The EM Algorithm is employed to obtain the maximum likelihood estimates of population parameters under different hypotheses. A parametric test procedure for testing the equality of means of two independent log-normal in the presence of censored data with single detection limit is presented. The performance of the test procedure presented in this article is evaluated by means of simulation studies. A detailed case study example of the method is provided using copper and zinc data presented in Millard and Deverel (1998). It is seen in the analysis of the Millard and Deverel (1998) data as shown in the study case example that large (unusual) data values do influence the estimate of the mean, but do not influence the estimate of the median in log-normal parametric model analysis. The nonparametric median comparison methods are not as sensitive to these unusual data values. I hope that my paper would be useful to researchers using the log-normal distribution in their analysis of the censored data.

Acknowledgements

The author is deeply indebted to the editor Professor Wendy Smith and the referees for their useful comments and recommendations which enhanced the clarity of the results of this work.

References

Aboueissa, A. A., & Stoline, R. M. (2004). Estimation of the Mean and Standard Deviation from Normally Distributed Singly-Censored Samples. *Environmetrics, 15,* 659-673. http://dx.doi.org/10.1002/env.643

Cohen, A. C. R. (1959). Simplified Estimators For The Normal Distribution When Samples Are Singly Censored Or Truncated. *Technometrics, 3,* 217-237. http://dx.doi.org/10.1080/00401706.1959.10489859

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *The Journal Of Royal Statistical Society B, 39*, 1-38.

El-Shaarawi, A. H. (1989). Inferences about the Mean from Censored Water Quality Data. *Water Resources Research, 25*, 685-690. http://dx.doi.org/10.1029/WR025i004p00685

El-Shaarawi, A. H., & Dolan, D. M. (1989). Maximum Likelihood Estimation Of Water Concentrations From Cen-

sored Data. *Canadian Journal Of Fisheries And Aquatic Sciences, 46,* 1033-1039. http://dx.doi.org/10.1139/f89-134

El-Shaarawi, A. H., & Esterby, S. R. (1992). Replacement Of Censored Observations By A Constant: An Evaluation. *Water Research, 26*(6), 835-844. http://dx.doi.org/10.1016/0043-1354(92)90015-V

Gibbons, R. D. (1994). *Statistical Methods For Groundwater Monitoring*, John Wiley & Sons, New York. http://dx.doi.org/10.1002/9780470172940

Gilbert, R. O. (1987). *Statistical Methods For Environmental Pollution Monitoring*, Van Nostrand Reinhold: New York.

Gleit, A. (1985). Estimation for small normal data sets with detection limits. *Environ. Sci. Technol., 19,* 1201-1206. http://dx.doi.org/10.1021/es00142a011

Gupta, R. C., & Li, X. (2006). Statistical Inference for the Common Mean of two Log-normal Distributions and some Applications in Reliability. *Computational Statistics and Data Analysis, 50,* 3141-3164. http://dx.doi.org/10.1016/j.csda.2005.05.005

Harris, G. A. (1991). Two-samples Comparisons in the Presence of Less-than-detectable data. *Proceeding of the Section on Statistics and the Environment: American Statistical Association*, 197-201.

Marco, B. (2005). *On Maximum Likelihood Estimation of Operational Loss Distributions*: Universita Degli, Studi Di Trento, Discussion paper No. 3.

Millard, S. P., & Deverel, S. J. (1998). Nonparametric Statistical Methods for Comparing two Sites Based on Data with multiple Nondetect Limits. *Water Resources Research, 24*, 2087 - 2098. http://dx.doi.org/10.1029/WR024i012p02087

Prentice, R. L. (1978). Linear rank Tests with Right Censored Data. *Biometrika, 65,* 167-179. http://dx.doi.org/10.1093/biomet/65.1.167

Schneider, H. (1986). *Truncated and Censored Samples from Normal Population*. Marcel Dekker: New York.

Stoline, M. R. (1993). Comparison of Two Medians Using a Two-Sample Log-normal Model in Environmental Contexts. *Environmetrics, 4*(3), 323-339. http://dx.doi.org/10.1002/env.3170040307

USEPA. (1989b). *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Interim Final Guidance. EPA/530-SW-89-026. Office of Solid Waste*, U.S. Environmental Protection Agency: Washington, D.C.

Wolynetz, M. S. (1979). Maximum Likelihood Estimations from Confined and Censored Normal Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(2), 185 - 195.

Yan, J., Misty, J. H., James, A. D., & Cynthia, J. H. (2010). Analysis of Lognormally Distributed Exposure Data with Repeated Measures and Values below the Limit of Detection Using SAS. Oxford University Press. *British Occupational Hygiene Society, 55*(1), 97-112.

Zhou, X., Sujuan, G., & Hui, S. L. (1997). Methods for Comparing the Means of Two Independent Log-normal Samples. *Biometrics, 53,* 1129-1135. http://dx.doi.org/10.2307/2533570

Appendix

Computer Programs

The following computer program, "*Abou.Two.Lognormal.Estimates*", is written in the R language to automate parameters estimation from left-censored data sets that are normally or log-normally distributed and to obtain the estimated values of the log-likelihood functions under the hypotheses H_{0N} and H_{A_1N}. In addition, this computer program will be used to obtain the asymptotic $\alpha-$level chi-square test statistic and its p-value.

```
Abou.Two.Lognormal.Estimates<-function(data1, data2, NumI, LogN) {
     #
     # NumI is the number of iterations suggested by users.
     # data1 and data2 are matrices containing of two columns each
     # the first column is the data set and the second column
```

```
     # is indicator 0 for uncensored and 1 for censored observations.
     # LogN = T if the data are log-normally distributed

    n1<-length(data1[,1])
    n2<-length(data2[,1])

    table1 <- table(data1[data1[, 2]==1, 1])
    DLV1<-as.numeric(dimnames(table1)[[1]])
    mcV1<-as.vector(table1)
    ##print(mcV1)
    DL1<-mean(DLV1)
    table2 <- table(data2[data2[, 2]==1, 1])
    DLV2<-as.numeric(dimnames(table2)[[1]])
    mcV2<-as.vector(table2)
    DL2<-mean(DLV2)
    for(i in 1:n1){
    if(data1[i,2]==1) data1[i,1]<-DL1 else data1[i,1]<-data1[i,1]
    }
    for(i in 1:n2){
    if(data2[i,2]==1) data2[i,1]<-DL2 else data2[i,1]<-data2[i,1]
    }
    if(LogN==T) data1[,1]<-log(data1[,1]) else data1[,1]<-data1[,1]

    if(LogN==T) data2[,1]<-log(data2[,1]) else data2[,1]<-data2[,1]
    datacomb<-rbind(data1,data2)
    n<-length(datacomb[,1])
    table <- table(datacomb[datacomb[, 2]==1, 1])
    DLV<-as.numeric(dimnames(table)[[1]])
    mcV<-as.vector(table)
    k<-length(mcV)
##################   EM Algorithm    ##################
AbouEMmultvect<-function(data, NumI) {
     #
     # N is the number of iterations suggested by users.
     # data is a matrix containing of two columns
     # the first column is the data set and the second column
     # is indicator 0 for uncensored and 1 for censored obs.
     #
    n<-length(data[,1])
    table <- table(data[data[, 2]==1, 1])
    DLV<-as.numeric(dimnames(table)[[1]])
    mcV<-as.vector(table)
    Xmbar<-tapply(data[,1],list(data[,2]),mean)["0"]
    Smsquare<-tapply(data[,1],list(data[,2]),var)["0"]
    g<-Smsquare/(Xmbar-(sum(DLV)/2))^2
    n<-length(data[,1])
    m<-sum(data[,2]==0)
    k<-length(DLV)
    mc<-numeric(k)
        mc<-numeric(k)
    u<-numeric(n)
        for(r in 1:k) {
    for(i in 1:n) {
    if(data[i,1]==DLV[r] && data[i,2]==1)
    u[i]<-1
    else
    u[i]<-0
    }
    mc[r]<-sum(u)
    }
    mu0.hat<-Xmbar
    sig0.hat<-Smsquare

    muhat<-numeric(NumI)
    sighat<-numeric(NumI)
        w<-matrix(0,n,2)
    ww<-matrix(0,n,2)
    w[,2]<-data[,2]
    ww[,2]<-data[,2]
        for(i in 1:n) {
        if(data[i,2]==1) {
        z0<-(data[i,1]-mu0.hat)/sqrt(sig0.hat)
        d0<-dnorm(z0)
        p0<-pnorm(z0)
          wdp0<-d0/p0
          w[i,1]<-mu0.hat-(sqrt((sig0.hat))*wdp0)
          ww[i,1]<-(wdp0)*(wdp0+z0)
                }
           else {
        w[i,1]<-data[i,1]
```

```
        ww[i,1]<-data[i,1]
        }
        muhat[1]<-mean(w[,1])
        num0<-sum((w[,1]-muhat[1])^2)
        dnum1<-tapply(ww[,1],list(ww[,2]),sum)["1"]
        dnum0<-m+dnum1
        sighat[1]<-num0/dnum0
        }
        for(j in 2:NumI) {
        for(i in 1:n) {
        if(data[i,2]==1) {
        ze<-(data[i,1]-muhat[j-1])/sqrt(sighat[j-1])
        de<-dnorm(ze)
        pe<-pnorm(ze)
          wdpe<-de/pe
          w[i,1]<-muhat[j-1]-(sqrt((sighat[j-1]))*wdpe)
          ww[i,1]<-(wdpe)*(wdpe+ze)
        }
        else {
        w[i,1]<-data[i,1]
        ww[i,1]<-data[i,1]
            }
        muhat[j]<-mean(w[,1])
        nume<-sum((w[,1]-muhat[j])^2)
        dnum2<-tapply(ww[,1],list(ww[,2]),sum)["1"]
        dnume<-m+dnum2
        sighat[j]<-nume/dnume

        }
    if(abs(muhat[j]-muhat[(j-1)])<1e-007 && abs(sighat[j]
    -sighat[(j-1)])<1e-007) break
    muhatf<-muhat[j]
        sigsqhatf<-sighat[j]
    sighatf<-sqrt(sighat[j])

                }
    musighat<-c(muhatf,sighatf)
    musighat

    }

EM.EstimatesPooled<-AbouEMmultvect(datacomb,NumI)
EM.Estimates1<-AbouEMmultvect(data1,NumI)
EM.Estimates2<-AbouEMmultvect(data2,NumI)
EM.Estimates<-rbind(EM.EstimatesPooled,EM.Estimates1,EM.Estimates2)
datacombest<-numeric(n)
    for(i in 1:n){
    if(datacomb[i,2]==1) datacombest[i]<-log(pnorm((datacomb[i,1]
    -EM.EstimatesPooled[1])/EM.EstimatesPooled[2]))
        else datacombest[i]<-log((1/EM.EstimatesPooled[2])
    *dnorm((datacomb[i,1]-EM.EstimatesPooled[1])/EM.EstimatesPooled[2]))
    }
    Loglikelihood.H0<-sum(datacombest)
data1est1<-numeric(n1)
    for(i in 1:n1){
    if(data1[i,2]==1) data1est1[i]<-log(pnorm((data1[i,1]
    -EM.Estimates1[1])/EM.Estimates1[2]))
        else data1est1[i]<-log((1/EM.Estimates1[2])*dnorm((data1[i,1]
    -EM.Estimates1[1])/EM.Estimates1[2]))
    }
    Loglikelihood.HAdata1<-sum(data1est1)
data1est2<-numeric(n2)
    for(i in 1:n2){
    if(data2[i,2]==1) data1est2[i]<-log(pnorm((data2[i,1]
    -EM.Estimates2[1])/EM.Estimates2[2]))
        else data1est2[i]<-log((1/EM.Estimates2[2])
    *dnorm((data2[i,1]-EM.Estimates2[1])/EM.Estimates2[2]))
    }
    Loglikelihood.HAdata2<-sum(data1est2)
    Loglikelihood.HA<-Loglikelihood.HAdata1+Loglikelihood.HAdata2
    chisquare0<- -2*(Loglikelihood.H0 - Loglikelihood.HA)
    p.value<- 1 - pchisq(chisquare0 , 1)
    Test.Result <- c(chisquare0,p.value)
    Test.Output<- rbind(EM.EstimatesPooled,EM.Estimates1
    ,EM.Estimates2,Test.Result)
    Test.Output
As<-matrix(0,4,6)
    As[1,1]<-"------"
    As[1,2]<-"----------"
    As[1,3]<-"----------------"
```

```
    As[1,4]<-"----------------"
    As[1,5]<-"----------"
    As[1,6]<-"-------"
    As[2,1]<-round(EM.EstimatesPooled[1], 4)
    As[2,2]<-round(EM.EstimatesPooled[2], 4)
    As[2,3]<-round(Loglikelihood.H0, 4)
    As[2,4]<-round(Loglikelihood.HA ,4)
    As[2,5]<-round(chisquare0, 4)
    As[2,6]<-round(p.value, 4)
    As[3,1]<-round(EM.Estimates1[1], 4)
    As[3,2]<-round(EM.Estimates1[2], 4)
    As[3,3]<-"       "
    As[3,4]<-"       "
    As[3,5]<-"       "
    As[3,6]<-"       "
    As[4,1]<-round(EM.Estimates2[1], 4)
    As[4,2]<-round(EM.Estimates2[2], 4)
    As[4,3]<-"       "
    As[4,4]<-"       "
    As[4,5]<-"       "
    As[4,6]<-"       "
    dimnames(As)<-list(c( "    ", " Poold.Data:   ", "     Data 1:   "
    , "     Data 2:   ") , c("mu.hat","sigma.hate","loglikelihood.H0"
    ,"loglikelihood.HA", "Chisquare0", "P Value"))
    print(As,quote=F)
        invisible()
    }

ZincALL<-matrix(c(10,9,5,18,10,12,10,11,11,19,8,3,10,10,10,10,10,10,10,10,10
          ,10,10,10,10,10,10,20,20,10,20,20,20,10,10,20,620,40,50,33
          ,10,20,10,20,10,30,20,10,20,20,20,10,20,23,17,10,10,10,20
          ,29,20,10,10,10,10,7,10,1,0,0,0,1,0,0,0,0,0,0,1,1,1,0,0,0
          ,0,1,0,1,0,1,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
          ,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,1,0,0,1),67,2)
ZincBas<-matrix(c(20,10,60,20,12,8,10,14,10,17,3,11,5,12,4,3,6,3,15,3,4,20,20
          ,70,60,40,30,40,17,10,20,20,5,10,50,30,25,10,10,40,20,10,20
          ,20,30,20,30,50,90,20,0,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,0
          ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
          ,0),50,2)

Abou.Two.Lognormal.Estimates(ZincALL, ZincBas, 20, T)


> Abou.Two.Lognormal.Estimates(ZincALL, ZincBas, 20, T)

              mu.hat sigma.hate loglikelihood.H0 loglikelihood.HA Chisquare0 P Value
              ------ ---------- ---------------- ---------------- ---------- -------
 Poold.Data:  2.5241   0.91       -147.9623        -146.5961        2.7324   0.0983
     Data 1: 2.4048    0.8868
     Data 2: 2.6825    0.9181
```

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).