

Comparison of Means of Two Log-Normal Distributions when Data is Multiply Censored

Abou El-Makarim A. Aboueissa

Department of Mathematics and Statistics, University of Southern Maine, 96 Falmouth Street, P.O. Box 9300, Portland, Maine 04104-9300, USA

Received: May 6, 2015 Accepted: June 10, 2015 Online Published: July 10, 2015

doi:10.5539/ijsp.v4n3p127 URL: <http://dx.doi.org/10.5539/ijsp.v4n3p127>

Abstract

When measuring concentration of chemical compounds, we often have to deal with a situation when the resulting values are found below the detection limit of the determination method. In order to statistically evaluate such data, the newly developed method of maximum likelihood considering multiply left-censored samples is applied. This paper is motivated by the need to have valid inference concerning the equality of the means of two log-normal distributions that are frequently encountered in environmental and exposure data analysis. As a model distribution of measured environmental and/or biomedical data, log-normal distribution is considered. Moreover, using the asymptotic properties of maximum likelihood estimates, concentrations of chemicals can be compared. A test procedure for comparing the means of two independent log-normal populations in the presence of multiply 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 and the size of the proposed test procedure introduced in this article.

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

1. Introduction

Many environmental data sets are characterized by a small number of high concentrations and a large number of low concentrations and are often right-skewed (Shumway et al., 1989). The log-normal distribution is positively skewed and hence can incorporate the few unusually high measurements of such environmental data in its long right-hand tail. For this reason the log-normal distribution is often applied to environmental data (Gilbert, 1987). While analyzing environmental and exposure data, a very common phenomenon is the occurrence of non-detects, i.e., observations below an analytical detection limit (DL), resulting in Type I singly left censored samples. Detection limit is the lowest concentration level that can be determined to be statistically different from a blank. The presence of observations below the DL significantly complicates the data analysis. Faced with such data, several strategies have been recommended for data analysis. One approach consists of replacing the below DL values with a constant such as $\frac{DL}{2}$, and using methods available for complete samples. It is easy to demonstrate that the conclusions resulting from this routine practice can be seriously flawed; in fact, the conclusions may depend on the substitution value used for replacing sample values below the 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. A sample is multiply censored if there are several detection limits. When more than two distinct detection limits DL_1, DL_2, \dots, 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. Multiple censoring commonly occurs with environmental data because detection limits can change over time (e.g., because of analytical improvements), or detection limits can depend on the type of sample or the background matrix. 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. 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. Paul and Gary (2007) compare the performance of several methods for statistically analyzing censored data sets when estimating the 95th percentile and the mean of right-skewed occupational exposure data. Krishnamoorthy et al (2014) proposed tests and confidence intervals for the ratio of the two means of two log-normal distributions, based on pivotal quantities involving the maximum likelihood estimators. Other suggested methods for comparing the means of two log-normal distributions are discussed in Krishnamoorthy et al (2014, 2011, 2007, 2006, 2003). Some of these methods are based on the generalized p-value and generalized confidence intervals, and others are based on the generalized test variable. Aboueissa (2015) introduced a test procedure for comparing the means of two independent log-normal populations when data is singly censored. Abdollahnezhad et al (2012) introduced a new method of test for comparing the means of two log-normal populations through the generalized measure of evidence to have against the null hypothesis. 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 (2011), 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 develop a parametric procedure to test the hypothesis of equal means when data are sampled from two independent log-normal distributions utilizing multiply left-censored data sets. The EM algorithm will be used to obtain the maximum likelihood estimates of population parameters under different hypotheses. To facilitate the application of this procedure, a computer program is written in the R language which calculates the maximum likelihood estimates, and 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. A simulation study was performed to inspect the size and the power of the proposed test procedure.

2. Assumptions and Notations

Assume that there exists two random samples of n_1 and n_2 data values: $y_{11}, y_{12}, \dots, y_{1m_1}, y_{1m_1+1}, \dots, y_{1n_1}$ and $y_{21}, y_{22}, \dots, y_{2m_2}, y_{2m_2+1}, \dots, y_{2n_2}$ 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

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

where $-\infty < \mu < \infty$ and $\sigma > 0$. For convenience, for each sample i let us assume that the first m_i observations $y_{i1}, y_{i2}, \dots, y_{im_i}$ are non-censored (fully measured) and the remaining $m_{c_i} = n_i - m_i$ observations are left-censored for $i = 1, 2$. For left censored observations, it is assumed that for each sample i it is only known that $y_{ij} < LDL_{ij}$ for $j = m_i + 1, \dots, n_i$ (or $j = 1, 2, \dots, m_{c_i}$) and $i = 1, 2$. The parameters for the i^{th} log-normal population can be

expressed as functions of the parameters μ and σ as:

$$\begin{aligned} \text{mean:} & \quad \mu_{y_i} = e^{\mu_i + \frac{\sigma_i^2}{2}} \\ \text{median:} & \quad M_{y_i} = e^{\mu_i} \\ \text{variance:} & \quad \sigma_{y_i}^2 = \gamma_i(\gamma_i - 1) e^{2\mu_i} \\ \text{skewness:} & \quad s_{y_i} = (\gamma_i + 2) \sqrt{\gamma_i - 1} \end{aligned}$$

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

Let

$$x_{ij} = \begin{cases} \log(y_{ij}), & \text{for } i = 1, 2 \text{ and } j = 1, 2, \dots, m_i, \\ DL_{ij} = \log(DDL_{ij}), & \text{for } i = 1, 2 \text{ and } j = m_i + 1, m_i + 2, \dots, n_i. \end{cases}$$

where DDL_{ij} are the detection limits in the i^{th} log-normal sample and $m_i + m_{c_i} = n_i$ for $i = 1, 2$ and $j = m_i + 1, m_i + 2, \dots, n_i$.

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 $x_{ij} = \log(y_{ij})$ 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

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

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

$$\Phi(\xi_{ij}) = \int_{-\infty}^{\xi_{ij}} \phi(t) dt, \quad \text{where} \quad \xi_{ij} = \frac{DL_{ij} - \mu_i}{\sigma_i},$$

for $i = 1, 2$ and $j = m_i + 1, m_i + 2, \dots, n_i$.

We also define

$$W(x) = \frac{\phi(x)}{\Phi(x)} \quad \text{and} \quad z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad \text{for } i = 1, 2 \text{ and } j = 1, 2, \dots, m_i.$$

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

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1, \sigma_2) &= \prod_{i=1}^2 \left(\prod_{j=1}^{m_i} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2} \prod_{j=m_i+1}^{n_i} P(x_{ij} < DL_{ij}) \right) \\ &= \prod_{i=1}^2 \left(\prod_{j=1}^{m_i} \frac{1}{\sigma_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2} \prod_{j=m_i+1}^{n_i} P\left(\frac{x_{ij} - \mu_i}{\sigma_i} < \frac{DL_{ij} - \mu_i}{\sigma_i} \right) \right) \\ &= \prod_{i=1}^2 \left(\prod_{j=1}^{m_i} \frac{1}{\sigma_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2} \prod_{j=m_i+1}^{n_i} P\left(Z < \frac{DL_{ij} - \mu_i}{\sigma_i} \right) \right) \end{aligned}$$

which can be written as

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1, \sigma_2) &= \prod_{i=1}^2 \left[\prod_{j=1}^{m_i} \frac{1}{\sigma_i} \phi\left(\frac{x_{ij} - \mu_i}{\sigma_i} \right) \prod_{j=m_i+1}^{n_i} \Phi\left(\frac{DL_{ij} - \mu_i}{\sigma_i} \right) \right] \\ &= \prod_{i=1}^2 \left[\prod_{j=1}^{m_i} \frac{1}{\sigma_i} \phi(z_{ij}) \prod_{j=m_i+1}^{n_i} \Phi(\xi_{ij}) \right] \end{aligned} \tag{2.1}$$

where

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \text{ for } i = 1, 2 \text{ and } j = 1, \dots, m_i,$$

and

$$\xi_{ij} = \frac{DL_{ij} - \mu_i}{\sigma_i} \text{ for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i.$$

The two log-normal population means are confirmed equal whenever no evidence was available to reject the null hypothesis $H_{0LN} : \mu_{y_1} = \mu_{y_2}$ in favor of the alternative hypothesis $H_{ALN} : \mu_{y_1} \neq \mu_{y_2}$. Equivalently the two log-normal population means are confirmed equal whenever the null hypothesis

$$\underline{H_{0N}} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 = \sigma_2 = \sigma \text{ (overall homogeneity).}$$

is accepted in favor of one of the alternative hypotheses:

$$\underline{H_{A1N}} : \mu_1 \neq \mu_2 \text{ and } \sigma_1 \neq \sigma_2 \text{ (overall heterogeneity),}$$

$$\underline{H_{A2N}} : \mu_1 \neq \mu_2 \text{ and } \sigma_1 = \sigma_2 = \sigma \text{ (mean heterogeneity, variance homogeneity),}$$

or

$$\underline{H_{A3N}} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 \neq \sigma_2 \text{ (mean homogeneity, variance heterogeneity).}$$

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} , H_{A1N} , H_{A2N} and H_{A3N} . The derivations of these estimates are now described.

3.1 Maximum Likelihood Estimates under H_{0N}

Under the hypothesis H_{0N} , x_{ij} , for $i = 1, 2$ and $j = 1, 2, \dots, n_i$, 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, for each sample i let us assume that the first m_i observations $x_{i1}, x_{i2}, \dots, x_{im_i}$ are non-censored (fully measured) and the remaining $m_{c_i} = n_i - m_i$ observations are left-censored for $i = 1, 2$. For left censored observations, it is assumed that for each sample i it is only known that $x_{ij} < DL_{ij}$ for $j = m_i + 1, \dots, n_i$ (or $j = 1, 2, \dots, m_{c_i}$) and $i = 1, 2$.

The likelihood function $L_{H_{0N}}(\mu, \sigma)$ under H_{0N} is given by:

$$L_{H_{0N}}(\mu, \sigma) = \prod_{i=1}^2 \left(\sigma^{-m_i} \left(\frac{1}{\sqrt{2\pi}} \right)^{m_i} e^{-\frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma} \right)^2} \prod_{j=m_i+1}^{n_i} \Phi \left(\frac{DL_{ij} - \mu}{\sigma} \right) \right) \quad (3.1)$$

Hence, the corresponding log-likelihood function $\ell_{H_{0N}}(\mu, \sigma) = \log(L_{H_{0N}}(\mu, \sigma))$ of (3.1) is given by:

$$\begin{aligned} \ell_{H_{0N}}(\mu, \sigma) &= \sum_{i=1}^2 \left(-m_i \log \sigma + m_i \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma} \right)^2 \right) \\ &+ \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \left(\Phi \left(\frac{DL_{ij} - \mu}{\sigma} \right) \right) \end{aligned} \quad (3.2)$$

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

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

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

The maximum likelihood estimates $\hat{\mu}$ 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 :

$$\frac{\partial \ell_{h_{0N}}(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma} \right) - \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) = 0 \quad (3.3)$$

$$\frac{\partial \ell_{h_{0N}}(\mu, \sigma)}{\partial \sigma} = -m + \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma} \right)^2 - \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) \xi_{ij} = 0 \quad (3.4)$$

where $m = \sum_{i=1}^2 m_i$, $W(\xi_{ij}) = \frac{\phi(\xi_{ij})}{\Phi(\xi_{ij})}$ and $\xi_{ij} = \frac{DL_{ij} - \mu}{\sigma}$.

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:

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

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 $i = 1, 2; j = m_i + 1, 2, \dots, n_i$) is replaced by an estimate of $\hat{\mu}_s - \hat{\sigma}_s W\left(\frac{DL_{ij} - \hat{\mu}_s}{\hat{\sigma}_s}\right)$.

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

$$u_{ij} = \begin{cases} x_{ij}, & \text{for } i = 1, 2 \text{ and } j = 1, \dots, m_i \\ \hat{\mu}_s - \hat{\sigma}_s W\left(\frac{DL_{ij} - \hat{\mu}_s}{\hat{\sigma}_s}\right) & \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i \end{cases}$$

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

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

and

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

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.

3.2 Maximum Likelihood Estimates under H_{AIN}

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

$$L_{H_{AIN}}(\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^2 \left(\left(\frac{1}{\sqrt{2\pi}} \right)^{m_i} (\sigma_i)^{-m_i} e^{-\frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2} \prod_{j=m_i+1}^{n_i} \Phi \left(\frac{DL_{ij} - \mu_i}{\sigma_i} \right) \right) \quad (3.5)$$

Hence, the corresponding log-likelihood function $\ell_{H_{AIN}}(\mu_1, \mu_2, \sigma_1, \sigma_2)$ of (3.5) which is defined as $\log(L_{H_{AIN}}(\mu_1, \mu_2, \sigma_1, \sigma_2))$ is given by:

$$\begin{aligned} \ell_{H_{AIN}}(\mu_1, \mu_2, \sigma_1, \sigma_2) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \sigma_i \\ &\quad - \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \mu_i}{\sigma_i} \right) \end{aligned} \quad (3.6)$$

The maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i are the solutions to equations (3.7) and (3.8) for $i = 1, 2$.

$$\frac{\partial \ell_{H_{AIN}}(\mu_i, \sigma_i)}{\partial \mu_i} = \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right) - \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) = 0 \quad (3.7)$$

$$\frac{\partial \ell_{H_{AIN}}(\mu_i, \sigma_i)}{\partial \sigma_i} = \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma_i} \right)^2 - m_i - \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) \xi_{ij} = 0 \quad (3.8)$$

where $W(\xi_{ij}) = \frac{\phi(DL_{ij})}{\Phi(DL_{ij})}$ and $\xi_{ij} = \frac{DL_{ij} - \mu_i}{\sigma_i}$ for $i = 1, 2$.

The single sample EM algorithm estimation method can be used to obtain the maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i for $i = 1, 2$ as follows. At step 0 of the EM algorithm all non-censored observations are used to calculate the initial estimates of μ_i and σ_i for $i = 1, 2$ as follows:

$$\hat{\mu}_{i0} = \bar{x}_{m_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \quad \text{and} \quad \hat{\sigma}_{i0}^2 = s_{m_i}^2 = \frac{1}{m_i} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_{m_i})^2$$

For $i = 1, 2$ let $\hat{\mu}_{is}$ and $\hat{\sigma}_{is}$ be the maximum likelihood estimates of μ_i and σ_i at step s of this procedure. At step $s + 1$, each censored observation x_{ij} (where $i = 1, 2$; and $j = m_i + 1, \dots, n_i$) is replaced by an estimate of $\hat{\mu}_{is} - \hat{\sigma}_{is} W\left(\frac{DL_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_{is}}\right)$.

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

$$v_{ij} = \begin{cases} x_{ij}, & \text{for } i = 1, 2 \text{ and } j = 1, \dots, m_i \\ \hat{\mu}_{is} - \hat{\sigma}_{is} W\left(\frac{DL_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_{is}}\right) & \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i \end{cases}$$

So for $i = 1, 2$ the updated estimates $\hat{\mu}_{is+1}$ and $\hat{\sigma}_{is+1}$ of μ_i and σ_i are given by

$$\hat{\mu}_{is+1} = \frac{\sum_{j=1}^{m_i} v_{ij} + \sum_{j=m_i+1}^{n_i} v_{ij}}{n_i}$$

and

$$\hat{\sigma}_{is+1}^2 = \frac{\sum_{j=1}^{m_i} (v_{ij} - \hat{\mu}_{is+1})^2 + \sum_{j=m_i+1}^{n_i} (v_{ij} - \hat{\mu}_{is+1})^2}{m_i + \sum_{j=m_i+1}^{n_i} \gamma\left(\frac{v_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_{is}}\right)}$$

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

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

Convergence is achieved if $|\hat{\mu}_{1s} - \hat{\mu}_{1s+1}| < 0.00001$, $|\hat{\mu}_{2s} - \hat{\mu}_{2s+1}| < 0.00001$, $|\hat{\sigma}_{1s} - \hat{\sigma}_{1s+1}| < 0.00001$ and $|\hat{\sigma}_{2s} - \hat{\sigma}_{2s+1}| < 0.00001$ occur. When these convergence criteria are met, the maximum likelihood estimates for μ_1, μ_2, σ_1 and σ_2 are then given by $\hat{\mu}_1 = \hat{\mu}_{1s}$, $\hat{\mu}_2 = \hat{\mu}_{2s}$, $\hat{\sigma}_1 = \hat{\sigma}_{1s}$ and $\hat{\sigma}_2 = \hat{\sigma}_{2s}$, respectively.

3.3 Maximum Likelihood Estimates under H_{A2N}

Under the hypothesis H_{A2N} x_{ij} are assumed to be normally distributed with mean μ_i and standard deviation σ , for $i = 1, 2$ and $j = 1, 2, \dots, n_i$. Thus the likelihood function under H_{A2N} is given by:

$$L_{H_{A2N}}(\mu_1, \mu_2, \sigma) = \prod_{i=1}^2 \left(\left(\frac{1}{\sqrt{2\pi}} \right)^{m_i} (\sigma)^{-m_i} e^{-\frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma} \right)^2} \prod_{j=m_i+1}^{n_i} \Phi \left(\frac{DL_{ij} - \mu_i}{\sigma} \right) \right) \quad (3.9)$$

Hence, the corresponding log-likelihood function $\ell_{H_{A2N}}(\mu_1, \mu_2, \sigma)$ of (3.9) which is defined as $\log(L_{H_{A2N}}(\mu_1, \mu_2, \sigma))$ is given by:

$$\begin{aligned} \ell_{H_{A2N}}(\mu_1, \mu_2, \sigma) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \sigma \\ &\quad - \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \mu_i}{\sigma} \right) \end{aligned} \quad (3.10)$$

The maximum likelihood estimates $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\sigma}$ of μ_1, μ_2 and σ are the solutions to equations (3.11)-(3.13).

$$\frac{\partial \ell_{H_{A2N}}(\mu_1, \mu_2, \sigma)}{\partial \mu_1} = \sum_{j=1}^{m_1} \left(\frac{x_{1j} - \mu_1}{\sigma} \right) - \sum_{j=m_1+1}^{n_1} W(\xi_{1j}) = 0 \quad (3.11)$$

$$\frac{\partial \ell_{H_{A2N}}(\mu_1, \mu_2, \sigma)}{\partial \mu_2} = \sum_{j=1}^{m_2} \left(\frac{x_{2j} - \mu_2}{\sigma} \right) - \sum_{j=m_2+1}^{n_2} W(\xi_{2j}) = 0 \quad (3.12)$$

$$\frac{\partial \ell_{H_{A2N}}(\mu_1, \mu_2, \sigma)}{\partial \sigma} = \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu_i}{\sigma} \right)^2 - m - \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) \xi_{ij} = 0 \quad (3.13)$$

where $m = m_1 + m_2$, $W(\xi_{ij}) = \frac{\phi(DL_{ij})}{\Phi(DL_{ij})}$ and $\xi_{ij} = \frac{DL_{ij} - \mu_i}{\sigma}$ for $i = 1, 2$ and $j = m_i + 1, 2, \dots, n_i$.

The expectation maximization (EM) algorithm will be used iteratively to obtain the solutions $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\sigma}$ to the maximum likelihood equations (3.11) and (3.13). At step 0 of the EM algorithm all non-censored observations are used to calculate the initial estimates of μ_1, μ_2 and σ as follows:

$$\hat{\mu}_{10} = \bar{x}_{m_1} = \frac{1}{m_1} \sum_{j=1}^{m_1} x_{1j} \quad , \quad \hat{\mu}_{20} = \bar{x}_{m_2} = \frac{1}{m_2} \sum_{j=1}^{m_2} x_{2j} \quad ,$$

and

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

where

$$\hat{\mu}_0 = \frac{m_1 \hat{\mu}_{10} + m_2 \hat{\mu}_{20}}{m_1 + m_2}$$

Let $\hat{\mu}_{1s}, \hat{\mu}_{2s}$ and $\hat{\sigma}_s$ be the maximum likelihood estimates of μ_1, μ_2 and σ at step s of this procedure. At step $s + 1$, each censored observation x_{ij} (where $i = 1, 2; j = m_i + 1, 2, \dots, n_i$) is replaced by an estimate of $\hat{\mu}_{is} - \hat{\sigma}_s W \left(\frac{DL_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_s} \right)$.

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

$$t_{ij} = \begin{cases} x_{ij} \quad , & \text{for } i = 1, 2 \text{ and } j = 1, \dots, m_i \\ \hat{\mu}_{is} - \hat{\sigma}_s W \left(\frac{DL_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_s} \right) & \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i \end{cases}$$

So the updated estimates $\hat{\mu}_{1s+1}$, $\hat{\mu}_{2s+1}$ and $\hat{\sigma}_{s+1}$ of μ_1 , μ_2 and σ are given by

$$\hat{\mu}_{1s+1} = \frac{\sum_{j=1}^{m_1} t_{1j}}{m_1}, \quad \hat{\mu}_{2s+1} = \frac{\sum_{j=1}^{m_2} t_{2j}}{m_2}$$

and

$$\hat{\sigma}_{s+1}^2 = \frac{\sum_{i=1}^2 \sum_{j=1}^{m_i} (t_{ij} - \hat{\mu}_{is+1})^2}{\sum_{i=1}^2 m_i + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \gamma\left(\frac{t_{ij} - \hat{\mu}_{is}}{\hat{\sigma}_s}\right)}$$

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

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

Convergence is achieved if $|\hat{\mu}_{1s} - \hat{\mu}_{1s+1}| < 0.00001$, $|\hat{\mu}_{2s} - \hat{\mu}_{2s+1}| < 0.00001$ and $|\hat{\sigma}_s - \hat{\sigma}_{s+1}| < 0.00001$ occur. When these convergence criteria are met, the maximum likelihood estimates for μ_1 , μ_2 and σ are then given by $\hat{\mu}_1 = \hat{\mu}_{1s}$, $\hat{\mu}_2 = \hat{\mu}_{2s}$ and $\hat{\sigma} = \hat{\sigma}_s$, respectively.

3.4 Maximum Likelihood Estimates under H_{A3N}

The maximum likelihood estimators under the hypothesis H_{A3N} are more complicated to derive than those obtained under H_{0N} , H_{A1N} and H_{A2N} . In this section only an outline for obtaining estimates of μ , σ_1 and σ_2 for the hypothesis under consideration will be given. Under the hypothesis H_{A3N} x_{ij} are assumed to be normally distributed with mean μ and standard deviation σ_i , for $i = 1, 2$ and $j = 1, 2, \dots, n_i$. Thus the likelihood function under H_{A3N} is given by:

$$L_{H_{A3N}}(\mu, \sigma_1, \sigma_2) = \prod_{i=1}^2 \left(\frac{1}{\sqrt{2\pi}} \right)^{m_i} (\sigma_i)^{-m_i} e^{-\frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma_i} \right)^2} \prod_{j=m_i+1}^{n_i} \Phi \left(\frac{DL_{ij} - \mu}{\sigma_i} \right) \quad (3.14)$$

Hence, the corresponding log-likelihood function $\ell_{H_{A3N}}(\mu, \sigma_1, \sigma_2)$ of (3.14) which is defined as $\log(L_{H_{A3N}}(\mu, \sigma_1, \sigma_2))$ is given by:

$$\begin{aligned} \ell_{H_{A3N}}(\mu, \sigma_1, \sigma_2) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \sigma_i \\ &\quad - \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma_i} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \mu}{\sigma_i} \right) \end{aligned} \quad (3.15)$$

The maximum likelihood estimates $\hat{\mu}$, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ of μ , σ_1 and σ_2 are the solutions to equations (3.16)-(3.18).

$$\frac{\partial \ell_{H_{A3N}}(\mu, \sigma_1, \sigma_2)}{\partial \mu} = \sum_{i=1}^2 \frac{1}{\sigma_i} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \mu}{\sigma_i} \right) - \sum_{i=1}^2 \frac{1}{\sigma_i} \sum_{j=m_i+1}^{n_i} W(\xi_{ij}) = 0 \quad (3.16)$$

$$\frac{\partial \ell_{H_{A3N}}(\mu, \sigma_1, \sigma_2)}{\partial \sigma_1} = \sum_{j=1}^{m_1} \left(\frac{x_{1j} - \mu}{\sigma_1} \right)^2 - m_1 - \sum_{j=m_1+1}^{n_1} W(\xi_{1j}) \xi_{1j} = 0 \quad (3.17)$$

$$\frac{\partial \ell_{H_{A3N}}(\mu, \sigma_1, \sigma_2)}{\partial \sigma_2} = \sum_{j=1}^{m_2} \left(\frac{x_{2j} - \mu}{\sigma_2} \right)^2 - m_2 - \sum_{j=m_2+1}^{n_2} W(\xi_{2j}) \xi_{2j} = 0 \quad (3.18)$$

where $W(\xi_{ij}) = \frac{\phi(DL_{ij})}{\Phi(DL_{ij})}$ and $\xi_{ij} = \frac{DL_{ij} - \mu}{\sigma_i}$ for $i = 1, 2$ and $j = m_i + 1, 2, \dots, n_i$.

The expectation maximization (EM) algorithm will be used iteratively to obtain the solutions $\hat{\mu}$, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ to the maximum likelihood equations (3.16)-(3.18). At step 0 of the EM algorithm all non-censored observations are used to calculate the initial estimates of μ , σ_1 and σ_2 as follows:

$$\hat{\mu}_0 = \bar{x}_m = \frac{\sum_{j=1}^{m_1} x_{1j} + \sum_{j=1}^{m_2} x_{2j}}{m_1 + m_2},$$

and

$$\hat{\sigma}_{10}^2 = s_{m_1}^2 = \frac{\sum_{j=1}^{m_1} (x_{1j} - \hat{\mu}_0)^2}{m_1} \quad \text{and} \quad \hat{\sigma}_{20}^2 = s_{m_2}^2 = \frac{\sum_{j=1}^{m_2} (x_{2j} - \hat{\mu}_0)^2}{m_2}.$$

Let $\hat{\mu}_s$, $\hat{\sigma}_{1s}$ and $\hat{\sigma}_{2s}$ be the maximum likelihood estimates of μ , σ_1 and σ_2 at step s of this procedure. At step $s + 1$, each censored observation x_{ij} (where $i = 1, 2$; $j = m_i + 1, 2, \dots, n_i$) is replaced by an estimate of $\hat{\mu}_s - \hat{\sigma}_{is} W\left(\frac{DL_{ij} - \hat{\mu}_s}{\hat{\sigma}_{is}}\right)$.

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

$$t_{ij} = \begin{cases} x_{ij}, & \text{for } i = 1, 2 \text{ and } j = 1, \dots, m_i \\ \hat{\mu}_s - \hat{\sigma}_{is} W\left(\frac{DL_{ij} - \hat{\mu}_s}{\hat{\sigma}_{is}}\right) & \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i \end{cases}$$

The solution of equations (3.16)-(3.18) for μ , σ_1 , and σ_2 is more complicated in this case than the cases H_{0N} , H_{A1N} and H_{A2N} since it involves solution of a cubic equation for μ . Thus at this point a procedure similar to the one used in cases H_{0N} , H_{A1N} and H_{A2N} is not valid. Let

$$\begin{aligned} B_1 &= \sum_{j=m_1+1}^{n_1} W\left(\frac{DL_{1j} - \hat{\mu}_s}{\hat{\sigma}_{1s}}\right) \left[W\left(\frac{DL_{1j} - \hat{\mu}_s}{\hat{\sigma}_{1s}}\right) + \frac{DL_{1j} - \hat{\mu}_s}{\hat{\sigma}_{1s}} \right] \\ &= \sum_{j=m_1+1}^{n_1} \gamma\left(\frac{DL_{1j} - \hat{\mu}_s}{\hat{\sigma}_{1s}}\right) = \sum_{j=m_1+1}^{n_1} \gamma(\xi_{1j}) \end{aligned} \quad (3.19)$$

$$\begin{aligned} B_2 &= \sum_{j=m_2+1}^{n_2} W\left(\frac{DL_{2j} - \hat{\mu}_s}{\hat{\sigma}_{2s}}\right) \left[W\left(\frac{DL_{2j} - \hat{\mu}_s}{\hat{\sigma}_{2s}}\right) + \frac{DL_{2j} - \hat{\mu}_s}{\hat{\sigma}_{2s}} \right] \\ &= \sum_{j=m_2+1}^{n_2} \gamma\left(\frac{DL_{2j} - \hat{\mu}_s}{\hat{\sigma}_{2s}}\right) = \sum_{j=m_2+1}^{n_2} \gamma(\xi_{2j}) \end{aligned}$$

where $W(t)$, $\gamma(t)$, ξ_{1j} and ξ_{2j} were previously defined.

Also define

$$\bar{t}_1 = \frac{\sum_{j=1}^{n_1} t_{1j}}{n_1} \quad \text{and} \quad \bar{t}_2 = \frac{\sum_{j=1}^{n_2} t_{2j}}{n_2},$$

be the sample mean of the updated values in the i^{th} sample for $i = 1, 2$.

The maximum likelihood equations (3.16)-(3.18) can be written as

$$\frac{\hat{\sigma}_{1(s+1)}^2}{\hat{\sigma}_{2(s+1)}^2} = -\frac{n_1(\bar{t}_1 - \hat{\mu}_{s+1})}{n_2(\bar{t}_2 - \hat{\mu}_{s+1})}, \quad (3.20)$$

$$\frac{\sum_{j=1}^{n_1} (t_{1j} - \hat{\mu}_{s+1})^2}{\hat{\sigma}_{1(s+1)}^2} = m_1 + B_1, \quad (3.21)$$

$$\frac{\sum_{j=1}^{n_2} (t_{2j} - \hat{\mu}_{s+1})^2}{\hat{\sigma}_{2(s+1)}^2} = m_2 + B_2, \quad (3.22)$$

Aside: For the derivation of equation (3.20). At step $s + 1$ of the EM algorithm we have

$$t_{ij} = \hat{\mu}_{s+1} - \hat{\sigma}_{i(s+1)} W\left(\frac{DL_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}}\right) \quad \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i$$

which leads to

$$-\frac{t_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}} = W\left(\frac{DL_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}}\right) \quad \text{for } i = 1, 2 \text{ and } j = m_i + 1, \dots, n_i$$

Thus at step $s + 1$ equation (3.16) can be written as:

$$\begin{aligned} \sum_{i=1}^2 \frac{1}{\hat{\sigma}_{i(s+1)}} \sum_{j=1}^{m_i} \left(\frac{t_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}} \right) + \sum_{i=1}^2 \frac{1}{\hat{\sigma}_{i(s+1)}} \sum_{j=m_i+1}^{n_i} \left(\frac{t_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}} \right) &= 0 \\ \sum_{i=1}^2 \frac{1}{\hat{\sigma}_{i(s+1)}} \sum_{j=1}^{n_i} \left(\frac{t_{ij} - \hat{\mu}_{s+1}}{\hat{\sigma}_{i(s+1)}} \right) &= 0 \\ \frac{1}{\hat{\sigma}_{1(s+1)}} \sum_{j=1}^{n_1} \left(\frac{t_{1j} - \hat{\mu}_{s+1}}{\hat{\sigma}_{1(s+1)}} \right) + \frac{1}{\hat{\sigma}_{2(s+1)}} \sum_{j=1}^{n_2} \left(\frac{t_{2j} - \hat{\mu}_{s+1}}{\hat{\sigma}_{2(s+1)}} \right) &= 0 \\ \frac{n_1(\bar{t}_1 - \hat{\mu}_{s+1})}{\hat{\sigma}_{1(s+1)}^2} + \frac{n_2(\bar{t}_2 - \hat{\mu}_{s+1})}{\hat{\sigma}_{2(s+1)}^2} &= 0 \\ \frac{\hat{\sigma}_{1(s+1)}^2}{\hat{\sigma}_{2(s+1)}^2} &= -\frac{n_1(\bar{t}_1 - \hat{\mu}_{s+1})}{n_2(\bar{t}_2 - \hat{\mu}_{s+1})} \end{aligned}$$

which is equation (3.20).

By eliminating the expression $\frac{\hat{\sigma}_{1(s+1)}^2}{\hat{\sigma}_{2(s+1)}^2}$ from equations (3.20)-(3.22) yields a cubic equation in $\hat{\mu}_{s+1}$ which can be solved to obtain the estimate $\hat{\mu}_{s+1}$ of μ . Estimates $\hat{\sigma}_{1(s+1)}^2$ and $\hat{\sigma}_{2(s+1)}^2$ of σ_1 and σ_2 can then be obtained from (3.21) and (3.22), respectively. For the sake of simplicity, the specific method(s) used in the solution of the cubic equation is not presented here, but can it can be derived and programmed. Convergence is achieved if $|\hat{\mu}_s - \hat{\mu}_{s+1}| < 0.00001$, $|\hat{\sigma}_{1s} - \hat{\sigma}_{1(s+1)}| < 0.00001$ and $|\hat{\sigma}_{2s} - \hat{\sigma}_{2(s+1)}| < 0.00001$ occur. When these convergence criteria are met, the maximum likelihood estimates for μ , σ_1 and σ_2 are then given by $\hat{\mu} = \hat{\mu}_s$, $\hat{\sigma}_1 = \hat{\sigma}_{1s}$ and $\hat{\sigma}_2 = \hat{\sigma}_{2s}$, respectively.

4. Asymptotic Chi-Square Test

The estimated log-likelihood functions $\hat{\ell}_{H_{0N}}$, $\hat{\ell}_{H_{A1N}}$, $\hat{\ell}_{H_{A2N}}$ and $\hat{\ell}_{H_{A3N}}$ under the hypotheses H_{0N} (overall homogeneity), H_{A1N} (overall heterogeneity), H_{A2N} (mean heterogeneity, variance homogeneity) and H_{A3N} (mean homogeneity, variance heterogeneity), respectively; are obtained by replacing population parameters by their maximum likelihood estimates. Therefore from (3.2), (3.6), (3.10) and (3.15) we get:

$$\begin{aligned} \hat{\ell}_{H_{0N}}(\hat{\mu}, \hat{\sigma}) &= \sum_{i=1}^2 \left(-m_i \log \hat{\sigma} + m_i \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2} \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \hat{\mu}}{\hat{\sigma}} \right)^2 \right) \\ &+ \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \left(\Phi \left(\frac{DL_{ij} - \hat{\mu}}{\hat{\sigma}} \right) \right), \end{aligned} \quad (4.1)$$

$$\begin{aligned} \hat{\ell}_{H_{A1N}}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \hat{\sigma}_i \\ &- \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \right), \end{aligned} \quad (4.2)$$

$$\begin{aligned} \hat{\ell}_{H_{A2N}}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \hat{\sigma} \\ &- \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \hat{\mu}_i}{\hat{\sigma}} \right), \end{aligned} \quad (4.3)$$

and

$$\begin{aligned} \hat{\ell}_{H_{A3N}}(\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2) &= \sum_{i=1}^2 -\frac{m_i}{2} \log(2\pi) + \sum_{i=1}^2 -m_i \log \hat{\sigma}_i \\ &- \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^{m_i} \left(\frac{x_{ij} - \hat{\mu}}{\hat{\sigma}_i} \right)^2 + \sum_{i=1}^2 \sum_{j=m_i+1}^{n_i} \log \Phi \left(\frac{DL_{ij} - \hat{\mu}}{\hat{\sigma}_i} \right). \end{aligned} \quad (4.4)$$

In general, the asymptotic α -level chi-square test used to test the null hypothesis $H_0 : \theta = 0$ versus the alternative hypothesis $H_a : \theta \neq 0$ is defined by

$$\chi_0^2 = -2(\hat{\ell}_{H_0}(\hat{\theta}_0) - \hat{\ell}_{H_{A1N}}(\hat{\theta}_a)) > \chi_{(\alpha, df)}^2,$$

where χ_0^2 has a chi-square distribution with degrees of freedom df , which is defined by the number of free parameters under the alternative hypothesis H_a minus the number of free parameters under the null hypothesis H_0 , and $\chi_{(\alpha, df)}^2$ is the upper α -point value obtained from the chi-square table with degrees of freedom df .

The asymptotic α -level chi-square test used to test the equality of the means of two independent log-normal populations is now described.

Test 1: Overall Homogeneity versus Overall Heterogeneity

$$H_{0N} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 = \sigma_2 = \sigma \text{ versus}$$

$$H_{A1N} : \mu_1 \neq \mu_2 \text{ and } \sigma_1 \neq \sigma_2$$

The asymptotic α -level chi-square test to test the null hypothesis H_{0N} versus the alternative hypotheses H_{A1N} is defined by:

$$\chi_{0A1}^2 = -2(\hat{\ell}_{H_{0N}} - \hat{\ell}_{H_{A1N}}) > \chi_{(\alpha, 2)}^2 \quad (4.5)$$

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

$$p - \text{value} = P(\chi_{(2)}^2 > \chi_{0A1}^2). \quad (4.6)$$

Thus in this case the null hypothesis that the means of two independent log-normal populations are equal will be rejected if $\chi_{0A1}^2 > \chi_{(\alpha, 2)}^2$ or equivalently if $p - \text{value} < \alpha$.

Test 2: Overall Homogeneity versus Mean Heterogeneity and Variance Homogeneity

$$H_{0N} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 = \sigma_2 = \sigma \text{ versus}$$

$$H_{A2N} : \mu_1 \neq \mu_2 \text{ and } \sigma_1 = \sigma_2 = \sigma$$

The asymptotic α -level chi-square test to test the null hypothesis H_{0N} versus the alternative hypotheses H_{A12N} is defined by:

$$\chi_{0A2}^2 = -2(\hat{\ell}_{H_{0N}} - \hat{\ell}_{H_{A2N}}) > \chi_{(\alpha, 1)}^2 \quad (4.7)$$

where $\chi_{(\alpha, 1)}^2$ is the upper α -point for a chi-square random variable with 1 degrees of freedom. The p-value of this test statistic is defined by:

$$p - \text{value} = P(\chi_{(1)}^2 > \chi_{0A2}^2). \quad (4.8)$$

Thus in this case the null hypothesis that the means of two independent log-normal populations are equal will be rejected if $\chi_{0A2}^2 > \chi_{(\alpha, 1)}^2$ or equivalently if $p - \text{value} < \alpha$.

Test 3: Overall Homogeneity versus Mean Homogeneity and Variance Heterogeneity

$$H_{0N} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 = \sigma_2 = \sigma$$

versus

$$H_{A3N} : \mu_1 = \mu_2 = \mu \text{ and } \sigma_1 \neq \sigma_2$$

The asymptotic α -level chi-square test to test the null hypothesis H_{0N} versus the alternative hypotheses H_{A3N} is defined by:

$$\chi_{0A3}^2 = -2(\hat{\ell}_{H_{0N}} - \hat{\ell}_{H_{A3N}}) > \chi_{(\alpha, 1)}^2 \quad (4.9)$$

where $\chi_{(\alpha, 1)}^2$ is the upper α -point for a chi-square random variable with 1 degrees of freedom. The p-value of this test statistic is defined by:

$$p - \text{value} = P(\chi_{(1)}^2 > \chi_{0A3}^2). \quad (4.10)$$

Thus in this case the null hypothesis that the means of two independent log-normal populations are equal will be rejected if $\chi_{0A3}^2 > \chi_{(\alpha, 1)}^2$ or equivalently if $p - \text{value} < \alpha$.

Computer Programs: To facilitate the application of the test procedure and parameter estimation method described in this article, a computer program called "Abou.Two.Lognormal.Estimation.MultipleDL" is written in

the R language to automate parameters estimation from multiply left-censored data sets that are normally or log-normally distributed and to obtain the estimated values of the log-likelihood functions under the null and the alternative hypotheses . In addition, this computer program will be used to obtain the asymptotic α -level chi-square test statistic and its p-value. A Copy of the source code is given in the Appendix section and is available upon request.

For the sake of simplicity, in the remaining part of this article Test 1 (H_{0N} versus H_{A1N}) will be considered. Test 2 and Test 3 can be easily programmed and computed.

5. Example: Copper and Zinc Data

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. These data sets contains seven distinct detection limits (LDL :

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

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. 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 $\hat{\sigma}_1 = 0.8035$ with the 620 value included and $\hat{\sigma}_1 = 0.5799$ 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 $\hat{\mu}_1 = 2.4854$ and $\hat{\mu}_1 = 2.4665$ for these two cases, respectively. The corresponding estimates of the log-normal median zinc concentration in the Alluvial Fan Zone are $\hat{M}_{y_1} = 12.006$ with the 620 value included and $\hat{M}_{y_1} = 11.781$ with the 620 value removed. The comparable estimates of the log-normal mean zinc concentration are $\hat{\mu}_{y_1} = 16.580$ with the 620 value included and $\hat{\mu}_{y_1} = 13.922$ with the 620 value removed. In addition, the

estimates of the log-normal standard deviation for zinc concentration in the Alluvial Fan Zone are $\hat{\sigma}_{y_1} = 15.792$ with the 620 value included and $\hat{\sigma}_{y_1} = 8.767$ 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

Hypothesis	Copper		Zinc		ZincW620R	
	Alluvial Fan Zone	Basin-Trough Zone	Alluvial Fan Zone	Basin-Trough Zone	Alluvial Fan Zone	Basin-Trough Zone
Estimations of Normal Parameters						
H_{0N}	$\hat{\mu}_0 = 0.9825$ $\hat{\sigma}_0 = 0.8627$		$\hat{\mu}_0 = 2.5699$ $\hat{\sigma}_0 = 0.8641$		$\hat{\mu}_0 = 2.5478$ $\hat{\sigma}_0 = 0.7721$	
H_{A1N}	$\hat{\mu}_1 = 0.9442$ $\hat{\sigma}_1 = 0.8005$	$\hat{\mu}_2 = 1.0331$ $\hat{\sigma}_2 = 0.9355$	$\hat{\mu}_1 = 2.4854$ $\hat{\sigma}_1 = 0.8035$	$\hat{\mu}_2 = 2.6893$ $\hat{\sigma}_2 = 0.9174$	$\hat{\mu}_1 = 2.4665$ $\hat{\sigma}_1 = 0.5799$	$\hat{\mu}_2 = 2.6893$ $\hat{\sigma}_2 = 0.9174$
Estimations of Log-normal Parameters						
H_{0LN}	$\hat{\mu}_{0y} = 3.875$ $\hat{M}_{0y} = 2.671$ $\hat{\sigma}_{0y} = 4.073$		$\hat{\mu}_{0y} = 18.977$ $\hat{M}_{0y} = 13.065$ $\hat{\sigma}_{0y} = 19.993$		$\hat{\mu}_{0y} = 17.217$ $\hat{M}_{0y} = 12.779$ $\hat{\sigma}_{0y} = 15.544$	
H_{ALN}	$\hat{\mu}_{y1} = 3.542$ $\hat{M}_{y1} = 2.571$ $\hat{\sigma}_{y1} = 3.356$	$\hat{\mu}_{y2} = 4.352$ $\hat{M}_{y2} = 2.810$ $\hat{\sigma}_{y2} = 5.148$	$\hat{\mu}_{y1} = 16.580$ $\hat{M}_{y1} = 12.006$ $\hat{\sigma}_{y1} = 15.792$	$\hat{\mu}_{y2} = 22.424$ $\hat{M}_{y2} = 14.721$ $\hat{\sigma}_{y2} = 25.764$	$\hat{\mu}_{y1} = 13.922$ $\hat{M}_{y1} = 11.781$ $\hat{\sigma}_{y1} = 8.767$	$\hat{\mu}_{y2} = 22.424$ $\hat{M}_{y2} = 14.721$ $\hat{\sigma}_{y2} = 25.764$
Test	The Asymptotic Chi-square Test: χ_0^2 (P-value)					
H_{0N} vs. H_{A1N}	1.4504 (0.2285)		2.5845 (0.1079)		12.7312 (0.0004)	
Test	The P-value of the Median equality test, Millard and Deverel (1988)					
$m_{y1} = m_{y2}$	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 0.2285. 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 *NS2P* 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 0.1079 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 0.0004 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 *NS2P* 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, $n_1 = n_2 = 25$ 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 in each simulated sample. 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 levels. Censoring levels were set at the 15th, 30th, and 50th percentiles of the parent distribution(s) as shown in Tables 3 and 4. 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% , 30%	0.0692
Small ($n = 25$)	30% , 30%	0.0657
Large ($n = 75$)	15% , 30%	0.0586
Large ($n = 75$)	30% , 50%	0.0507

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 (0.0692, 0.0657) for the small sample size case, and slightly higher than 0.05 (0.0586, 0.0507) 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

$(\mu_1, \mu_2, \sigma_1, \sigma_2)$	Small Sample Size ($n = 25$)		Large Sample Size ($n = 75$)	
	Censoring Levels (15% , 30%)	Censoring Levels (30% , 50%)	Censoring Levels (15% , 30%)	Censoring Levels (30% , 50%)
(0, 0.1, 1, 1.1)	0.1027	0.0747	0.1827	0.1138
(0, 0.2, 1, 1.2)	0.2165	0.1538	0.4379	0.3556
(0, 0.3, 1, 1.3)	0.3318	0.2684	0.7434	0.6559
(0, 0.4, 1, 1.4)	0.4708	0.4279	0.9236	0.8783
(0, 0.5, 1, 1.5)	0.6393	0.5763	0.9817	0.9598
(0, 0.6, 1, 1.6)	0.7673	0.7024	0.9973	0.9872
(0, 0.7, 1, 1.7)	0.8505	0.8239	0.9996	0.9994
(0, 0.8, 1, 1.8)	0.9147	0.8680	1.0000	1.0000
(0, 0.9, 1, 1.9)	0.9435	0.9168	1.0000	1.0000
(0, 1.0, 1, 2.0)	0.9703	0.9425	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 $h_1 = 15\%$ and $h_2 = 30\%$ ($h_1 = 30\%$ and $h_2 = 50\%$) censoring level we reach a power of 0.9703 (0.9425) 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 $h_1 = 15\%$ and $h_2 = 30\%$ ($h_1 = 30\%$ and $h_2 = 50\%$) 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 0.7; and a power of 1.0 when the difference between μ 's and σ 's is 0.8.

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 levels moves from 0.15 and 0.30 to 0.30 and 0.50. 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 modelling 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 multiply 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 multiple 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 the researchers who are considering log-normal distribution in their analysis of the multiply left 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

- Abdollahnezhad, K., Babanezhad, M., & Jafari, A. A. (2012). Inference on Difference of Means of two Log-Normal Distributions; A Generalized Approach. *Journal of Statistical and Econometric Methods*, 1(2), 125-131.
- Aboueissa, A. A. (2015). Comparison of Two Means of Two Log-Normal Distributions When Data is Singly Censored. *International Journal of Statistics and Probability*, 4(2), 73-86.
- Cohen, A. C. R. (1959). Simplified Estimators For The Normal Distribution When Samples Are Singly Censored Or Truncated. *Technometrics*, 3, 217-237.
- Dempster, A. P., N. Laird, 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.
- El-Shaarawi, A. H., & Dolan, D. M. (1989). Maximum Likelihood Estimation Of Water Concentrations From Censored Data. *Canadian Journal Of Fisheries And Aquatic Sciences*, 46, 1033-1039.
- El-Shaarawi, A. H., & Esterby, S. R. (1992). Replacement Of Censored Observations By A Constant: An Evaluation. *Water Research*, 26(6), 835-844.
- 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.
- 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.
- Jianrong, W., Jiang, G., Wong, A., & Xiang, S. (2002). Likelihood Analysis for the Ratio of Means of Two Independent Log-Normal Distributions. *Biometrics*, 58, 463-469.
- Krishnamoorthy, K., Mathew, T., & Xu, Z. (2014). Comparison of Means of Two Lognormal Distributions Based on Samples with Multiple Detection Limits. *Journal of Occupational and Environmental Hygiene*, 11(8), 538-546.
- 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(5), 1-18.
- Krishnamoorthy, K., & Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized

- p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference*, 115, 103-121.
- Krishnamoorthy, K., Mathew, T., & Ramachandran, G. (2006). Generalized P-Values and Confidence Intervals: A Novel Approach for Analyzing Lognormally Distributed Exposure Data. *Journal of Occupational and Environmental Hygiene*, 3, 642-650.
- Krishnamoorthy, K., Avishek, M., & Mathew, T. (2011). Inference for the Lognormal Mean and Quantiles Based on Samples with Left and Right Type I Censoring. *Technometrics*, 53(1), 72-83.
- 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.
- 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
- Paul, H., & Gary, H. G. (2007). A Comparison of Several Methods for Analyzing Censored Data. *Oxford University Press on behalf of the British Occupational Hygiene Society*, 51(7), 611-632.
- Schneider, H. (1986). *Truncated and Censored Samples from Normal Population*. Marcel Dekker: New York.
- Shumway, R. H., Azari, A. S., & Johnson, P. (1989). Estimating mean concentrations under transformation for environmental data with detection limits. *Technometrics*, 31, 347-356.
- Stoline, M. R. (1993). Comparison of Two Medians Using a Two-Sample Log-normal Model in Environmental Contexts. *Environmetrics*, 4(3), 323-339.
- 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. (2011). 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.

Appendix Computer Programs

The following computer program, "Abou.Two.Lognormal.Estimation.MultipleDL", is written in the R language to automate parameters estimation from multiple 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_{A1N} . In addition, this computer program will be used to obtain the asymptotic α -level chi-square test statistic and its p-value.

```
Abou.Two.Lognormal.Estimation.MultipleDL<-function(data1, data2, NumI, LogN1, LogN2) { ### \\
#
# 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.
# NumI is the number of iterations suggested by users.
# LogN1 = T if the data1 are log-normally distributed.
# LogN2 = T if the data2 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)
table2<-table(data2[data2[,2]==1, 1])
DLV2<-as.numeric(dimnames(table2)[[1]])
mcV2<-as.vector(table2)
if(LogN1==T) data1[,1]<-log(data1[,1]) else data1[,1]<-data1[,1]
if(LogN2==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) {
#
```


