ORNL/TM-2005/52

Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-Detectable Values

SEPTEMBER 2005

Prepared by

E. L. Frome
Computer Science and Mathematics Division
Oak Ridge National Laboratory

Paul F. Wambach
U. S. Department of Energy

STATISTICAL METHODS AND SOFTWARE FOR THE ANALYSIS OF OCCUPATIONAL EXPOSURE DATA WITH NON-DETECTABLE VALUES

Edward L. Frome
Computer Science and Mathematics Division
Oak Ridge National Laboratory

Paul F. Wambach
U. S. Department of Energy

Date Published: September 2005

Prepared by
OAK RIDGE NATIONAL LABORATORY
P. O. Box 2008
Oak Ridge, Tennessee 37831-6285
managed by
UT-Battelle, LLC
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

ABSTRACT

Environmental exposure measurements are, in general, positive and may be subject to left censoring; i.e,. the measured value is less than a "detection limit." In occupational monitoring, strategies for assessing workplace exposures typically focus on the mean exposure level or the probability that any measurement exceeds a limit. Parametric methods used to determine acceptable levels of exposure, are often based on a two parameter lognormal distribution. The mean exposure level, an upper percentile, and the exceedance fraction are used to characterize exposure levels, and confidence limits are used to describe the uncertainty in these estimates. Statistical methods for random samples (without non-detects) from the lognormal distribution are well known for each of these situations. In this report, methods for estimating these quantities based on the maximum likelihood method for randomly left censored lognormal data are described and graphical methods are used to evaluate the lognormal assumption. If the lognormal model is in doubt and an alternative distribution for the exposure profile of a similar exposure group is not available, then nonparametric methods for left censored data are used. The mean exposure level, along with the upper confidence limit, is obtained using the product limit estimate, and the upper confidence limit on an upper percentile (i.e., the upper tolerance limit) is obtained using a nonparametric approach.

All of these methods are well known but computational complexity has limited their use in routine data analysis with left censored data. The recent development of the R environment for statistical data analysis and graphics has greatly enhanced the availability of high-quality nonproprietary (open source) software that serves as the basis for implementing the methods in this paper. Numerical examples are provided and R(2004) functions are available at the analysis of occupational exposure data web site **http://www.csm.ornl.gov/esh/aoed/** (AOED).

Key words: exposure measurements, lognormal, maximum likelihood, left censored, non-detect, confidence limits, tolerance limit, exceedance fraction, occupational monitoring.

INTRODUCTION

Regulatory and advisory criteria for evaluating the adequacy of occupational exposure controls are generally expressed as limits that are not to be exceeded in a work shift or shorter time-period if the agent is acutely hazardous. Exposure monitoring results above the limit require minimal interpretation and should trigger immediate corrective action. Demonstrating compliance with a limit is more difficult. The American Industrial Hygiene Association (AIHA) has published a consensus standard with two basic strategies for evaluating an exposure profile [Mulhausen and Damiano (1998)]. The first approach is based on the mean of the exposure distribution, and the second approach considers the "upper tail" of the exposure profile. Statistical methods for estimating the mean, an upper percentile of the distribution, the exceedance fraction, and the uncertainty in each of these parameters are described. Most of the AIHA methods are based on the assumptions that the exposure data does not contain non-detects, and that a lognormal distribution can be used to describe the data. Exposure monitoring results from a compliant workplace tend to contain a high percentage of non-detected results when the detection limit is close to the exposure limit, and in some situations, the lognormal assumption may not be reasonable. There are parametric methods for censored lognormal data and non-parametric methods that can be used with left censored data to calculate all of the statistics recommended by the AIHA for the complete data case. However, the only practical way to compute these statistics is with statistical software. The recent availability of free, high-quality statistical software means that complex calculations are no longer a barrier to the statistical analysis of almost any occupational exposure data set. This also eliminates the need for special tables and graphical methods that are used in the complete data case for the lognormal distribution.

Statistical methods for the analysis of right censored data using various parametric and non-parametric methods are well known and generally referred to as "survival analysis" [Cox and Oakes (1984) or Kabfleish and Prentice (1980)]. In this situation, the dependent or response variable (say T) is usually time to the occurrence of event; i.e., the "survival time" (or time to failure) of an observational or experimental unit (e.g., animal, person, or machine). T may be referred to as a "lifetime random variable" and is by definition positive, and may be subject to "censoring." As a typical example, let T_i represent the survival time of the *i*th patient in a clinical trial. If the trial ends and the patient is not known to have "failed" the observed survival time, say t_i^*, is right censored (i.e., it is only known that T_i is greater than t_i^*). This can occur for several reasons. Suppose, for example, that all patients enter the trial at the same time and are followed until a specified end date, then those individuals still at risk have a censored survival time that is the same for all surviving patients (Type I censoring). If patients enter the trial at random and the trial ends at a fixed date, then the value of t_i^* is different for each surviving patient (random censoring). Statistical methods for the analysis of right censored data are widely used and computer software for survival analyses is available in most general purpose statistical programs (e.g., the R survival library).

In this report, the dependent or response variable of interest is the amount, say X, of a measured quantity. X is a positive random variable and as the result of the analytic methods used, the observed value for the *i*th measurement may be reported as (left) "censored" and is referred to as a non-detect or as being less than a "detection limit"(DL) say x_i^* (i.e., it is only known that X_i is less than x_i^*). A frequent assumption is that the distribution of X is lognormal . See Aitchison and Brown(1969) and Crow and Shimizu(1988) for general treatment of the lognormal distribution and its application. Schmoyer et. al. (1996) considered the lognormal model for contaminant concentrations in environmental risk assessment for both complete and left censored samples. Akritas et. al. (1994) provide a detailed discussion of various methods that have been proposed for parameter estimation for left censored data. Methods were classified into three general areas as described by Helsel (1990). They are: (1) simple substitution; e.g., replace a censored observation with one-half of its value; (2) parametric methods; e.g., censored data maximum likelihood (ML), and (3) "robust parametric methods" based on variations of "probability plot

regression." Simulations studies have been done to compare these methods under various conditions with the primary focus on bias and mean square error of location and scale parameters [see Helsel and Cohen(1988), Newman et. al. (1989)]. Akritas et. al. (1994) have reviewed these and other studies, and note that ML methods under the lognormal model provide expressions for the variance of the parameter estimates. This is important when an upper percentile, say Xp , of the exposure distribution is of interest. For example, the ML estimate of $\log(X_p)$ is a linear combination of the lognormal parameters, and the standard error of this quantity can be estimated from the ML parameter covariance matrix. Consequently, confidence limits for both X_p and the exceedance fraction can be obtained using the ML approach. Taylor et. al. (2001) have noted that regarding all non-detected values as censored outcomes from a lognormal distribution may not always be appropriate. If there is reason to believe that a non-negligible proportion of the non-detects are "true zero" exposures, then a censored lognormal mixture model (a zero-inflated lognormal model with censoring) should be considered. ML methods for estimation and hypothesis testing are described and the relationship between ML parameter estimates from the mixture model and those based on either a left truncated or censored lognormal model are described. Moulton and Halsey(1995) emphasize that it is also possible that non-detects may be from a second (possibly lognormal) distribution rather than a point mass at zero. Fowlkes (1979) has described methods for studying the mixture of two lognormal distributions, although he did not consider left censored data. In situations where the lognormal model (or some other distribution such as the gamma or Weibull) is not reasonable, a non-parametric approach can be used. All of the methods just described can be implemented using R, the ML method being the most difficult. It is also possible to develop procedures for all of these methods in several proprietary statistical programs that are available commercially.

In the discussion that follows the generic term "acceptable" refers to the situation where the distribution of the exposure measurement X satisfies a specified criteria indicating, for example, that the workplace is "safe," or that the surfaces of a survey unit are "clean." The term survey unit describes all or part of an entity (e.g., building, piece of equipment) that is being evaluated. The term "unacceptable" means that the distribution of the exposure measurements indicates that the workplace or survey unit is "contaminated" or "hazardous." The formal statistical procedures used to demonstrate that an exposure distribution is acceptable is to state a null hypothesis in the form $H_0 : \theta \geq L$, where θ represents a parameter of the exposure distribution (e.g., the mean, a percentile, or the exceedance fraction), and $\theta \geq L$ indicates that the exposure distribution is unacceptable. Then, based on a random sample from the exposure distribution, an estimate of θ and an upper confidence limit (UCL) with a specified confidence level, say γ, are calculated. If the 100γ%UCL is less than L, then the null hypothesis is rejected and the exposure distribution is acceptable. A Type I error occurs if H_0 is rejected when it is true (i.e., the X distribution is incorrectly considered to be acceptable). This will occur with a probability (type I error rate) that is less than or equal to $\alpha = (1 - \gamma)$, with $\alpha = 1 - \gamma$ when $\theta = L$. These and other related procedures are described in detail in an occupational exposure context by Mulhausen and Damiano (1998).

2. STATISTICAL ANALYSIS FOR COMPLETE SAMPLES

Lyles and Kupper (1996) have discussed strategies for the assessment of workplace exposures using time-weighed average (TWA) exposure measurements on a representative sample of workers as a typical example. The TWA measurements are considered to be a random sample from a lognormal distribution without censoring. They describe "exact" statistical methods for testing either (1) the null hypothesis that the mean exposure level for a similar exposure group is below a certain limit; i.e., the long-term average permissible exposure limit, or (2) that a specified percentile of the X distribution does not exceed a limit L. To review what is known for the complete data case suppose that x_i, i = 1, …, n is a random sample from a lognormal distribution with mean $\mu_x = \exp(\mu_y + \sigma_y^2 / 2)$, where μ_y and σ_y^2 are the corresponding

mean and variance of $y_i = \log(x_i)$. Let $\bar{y} = \Sigma_i\ y_i /n$ and $s_y^2 = \Sigma (y_i - \bar{y})^2/(n-1)$ where s_y^2 is the unbiased estimator for σ_y^2.

2.1 CONFIDENCE LIMITS FOR THE MEAN EXPOSURE LEVEL

A number of methods have been proposed for calculating confidence limits for μ_x [e.g., Armstrong (1992)]. For Land's (1972) exact method the 100γ% UCL is exp [$(\bar{y} + ½\ s_y^2 + C\ s_y/\sqrt{(n-1)}$], where C depends on s_y, n, and α and requires special tables. This is the "best" (i.e., uniformly most powerful unbiased test) for complete samples. The 100γ% lower confidence limit (LCL) is obtained in a similar way. The two one-sided limits can be combined to obtain an approximate confidence interval. It is also possible to obtain an exact two-sided confidence interval using Land's exact method. The "best estimate" of μ_x in complete samples is the minimum variance unbiased estimate [see Hewett and Ganser (1997) for details]. Equivalent optimal methods for randomly left censored data have not been developed. Two approximate confidence limits described by Land (1972) for the complete data case can be used for left censored data.

The first method is attributed to D.R. Cox and is based on calculating an estimate of $\varphi = \log(\mu_x) = \mu_y + \sigma_y^2/2$. For the complete data case the minimum variance unbiased estimate of φ is $\tilde{\varphi} = \bar{y} + s_y^2/2$, and the variance of $\tilde{\varphi}$ is $\text{var}(\tilde{\varphi}) = \text{var}(\bar{y}) + ¼\ \text{var}(s_y^2) = s_y^2/n + ½\ s_y^4/(n-1)$. The 100γ% UCL for μ_x is exp [$\tilde{\varphi} + t\ \text{var}(\tilde{\varphi})^{1/2}$], and the 100γ% LCL for μ_x is exp [$\tilde{\varphi} - t\ \text{var}(\tilde{\varphi})^{1/2}$], where $t = t(\gamma, n-1)$ is the 100γ percentage point of Student's t distribution on n-1 degrees of freedom [Land (1972) and Armstrong (1992)]. The point estimate of μ_x for this method is $\exp(\tilde{\varphi})$. These estimates can be viewed as "bias adjusted" ML estimates, since the ML estimate of φ is $\hat{\varphi} = \hat{\mu}_y + \hat{\sigma}_y^2/2$, and its variance is estimated as $\text{var}(\hat{\varphi}) = \hat{\sigma}_y^2/n + \hat{\sigma}_y^4/(2n)$ where $\hat{\sigma}_y^2 = \Sigma (y_i - \bar{y})^2/n$. The ML estimate of the (arithmetic) mean of X is $\hat{\mu}_x = \exp(\hat{\varphi})$. The estimate of the 100γ% UCL is exp [$\hat{\varphi} + t\ \text{var}(\hat{\varphi})^{1/2}$], and the estimate of the 100γ% LCL is exp [$\hat{\varphi} - t\ \text{var}(\hat{\varphi})^{1/2}$]. For left censored data ML estimates of the above quantities are not available in closed form, but can be obtained numerically (Cohen, 1991). The bias adjustment of variance terms described above could be applied to the censored data ML estimates so that results will reduce to the complete data case as the proportion of non-detects goes to zero. The second approximate method for confidence limits for μ_x is obtained by calculating the sample mean $\bar{x}$ as the point estimate of μ_x and the approximate lower and upper limits are $\bar{x} \pm t(\gamma, n-1)\ s_x/\sqrt{n}$, where $s_x^2 = \Sigma_i (x_i - \bar{x})^2/(n-1)$. The central limit theorem implies that this method should converge to the exact limit as n becomes large. For left censored data the product limit estimate (PLE) (Schmoyer et al, 1996) is used to obtain a non-parametric estimate of $\bar{x}$ and approximate confidence limits for μ_x (see Section 3.5).

2.2 CONFIDENCE LIMIT FOR THE PTH PERCENTILE

Let X_p denote the 100pth percentile of the lognormal distribution. The point estimate is $x_p = \exp(\bar{y} + z_p\ s_y)$ where z_p is the pth quantile of the standard normal distribution. An exact 100γ% upper confidence limit for the pth percentile is $UX(p, \gamma) = \exp(\bar{y} + K\ s_y)$ and is referred to as the upper tolerance limit. The value of K depends on n, p, and γ and is obtained from the 100 γ percentile of the noncentral t distribution with n-1 degrees of freedom and noncentrality parameter $-\sqrt{n}\ z_p$ [Lyles and Kupper (1996);

Johnson and Welch (1940)]. The null hypothesis of interest is H_0: $X_p \geq L_p$ where L_p is a specified limit (e.g., the occupational exposure limit). If $UX(p, \gamma) < L_p$ then H_0 is rejected indicating that the exposure level is acceptable (i.e, workplace is "safe" or object is "clean"). In this situation the probability is γ (we are 100γ% confident) that at least 100p% of the X values are less than $UX(p, \gamma)$ which is less than L_p. Throughout this report, reference to an R function is indicated by **bold face** font. The R function **extol(**n,p,gam) (see the Appendix) will return the one-sided tolerance factor K for any reasonable values of n, p, and γ. Further, **extol(**n,p,1-gam) will return the factor K´ proposed by Tuggle (1982) that can be used to calculate the exact 100γ% lower confidence limit for the pth percentile $LX(p, \gamma) = \exp(\bar{y} + K's_y)$. If $LX(p, \gamma) > L_p$ the probability is γ that at least 100(1-p)% of the Xs are above L_p. The one-sided tolerance bounds can be combined to obtain an approximate two-sided tolerance interval which is a confidence interval for X_p. Hahn and Meeker (1991) discuss the relationship between exact one and two-sided tolerance bounds, confidence intervals for population percentiles, and other types of statistical intervals. The factors K and K´ obtained using **extol** are found in their Table A.12 for selected values of n, p, and γ and in Mulhausen and Damiano(1998) Appendix Table VII..

2.3 CONFIDENCE LIMITS FOR EXCEEDANCE FRACTION

Let F_L represent the proportion of the Xs that exceed a given limit L_p. The null hypothesis is H_0: $F_L \geq F^0 = 1\text{-}p$; i.e., F_0 is the maximum proportion of the population that can exceed the limit L_p. The point estimate of F_L is $f = 1\text{-}N(u)$, where $u = [\log(L_p) - \bar{y}]/s_y$ and N(u) is the standard normal distribution function. If the 100γ% UCL for F_L, $Uf(L_p,\gamma)$, is less than F_0 than H_0 is rejected (i.e. the object or workplace is acceptable). This limit is obtained by first calculating the 100γ% lower confidence limit for u . It can be shown that $\sqrt{n}\,u$ follows the non-central t distribution with n-1 degrees of freedom and noncentrality parameter δ. A lower 100γ% confidence limit for u is obtained by solving **pt**($\sqrt{n}\,u$, n-1, dl) = γ for the noncentrality parameter dl where **pt** is the distribution function for the noncentral t distribution. The value of $ul = d1/\sqrt{n}$ is the LCL for u and the 100γ% UCL for F_L is $Uf(L_p,\gamma) = 1\text{-}N(ul)$, i.e. we are 100γ% confident that at most $Uf(L_p, \gamma)$ percent of the Xs are greater than L_p. The 100γ% lower confidence limit $Lf(L_p, \gamma)$ is obtained in a similar way. The R function **efcl(x**, γ, L,T) returns lognormal-based point and 100γ% lower and upper confidence limits for the exceedance fraction F_L (expressed as a percent) for complete samples (see the Appendix).

The relationship between the upper tolerance limit and the exceedance fraction is summarized as follows:

H_0: $X_p \geq L_p$ reject H_0 if $UX(p, \gamma) < L_p$,
H_0: $F_L \geq F_0 = 1\text{-}p$ reject H_0 if $Uf(L_p, \gamma) < F_0$. (1)

If, for example, p = .95, γ=0.95, and $L_p = 0.2$ then $F_0 = .05$ (5%). H_0 is rejected if $UX(.95,.95) \leq 0.2$, or if $Uf(0.2, 0.95) \leq 5\%$, and the exposure profile is considered acceptable. Note that if the lower confidence limit $Lf(L, \gamma) > F_0 = 1\text{-}p$ this indicates that, with confidence level gamma, at least $Lf(L, \gamma)$% of the Xs exceed L_p. This is equivalent to finding that the lower confidence limit for X_p is greater than L_p. Tuggle(1982) provides futher discussion of tolerance limits (and implicitly the exceedance fraction) as well as the choice of the values of p, L_p, and γ that determine conditions for an acceptable level of exposure. It is of interest to note that confidence limits for F require L_p and γ (not p), and confidence limits for X_p require p and γ (not L_p). The Type I error for this procedure is $\leq \alpha = (1\text{-}\gamma)$ = probability of rejecting Ho when it is true; i.e., incorrectly deciding that the exposure is acceptable (e.g., workplace is safe). The Type II error, β, is the probability of failing to reject H_0 when the alternative hypothesis, say $F_L = F^*$ ($< F_0$) is true, i.e. deciding that an acceptable exposure profile is not acceptable (clean object is contaminated.) The power = (1- β) of this test is the probability of correctly deciding that an exposure profile is acceptable. The power depends on the sample size and the "true" value of the exceedance fraction when the alternative hypothesis is true. For complete samples from the lognormal distribution

the power is **pt**(t_0 ,n-1,λ_1) where the non-centrality parameter is $\lambda_1 = z_{F*}\sqrt{n}$, and $t_0 = t'(n-1,\alpha,\lambda_0)$ is the 100αth percentile of the non-central t distribution with n-1 degrees of freedom and non-centrality parameter $\lambda_0 = z_F\sqrt{n}$. The exact sample size required to provide power of at least (1- β) may be obtained (Lyles and Kupper,1996) by finding the smallest integer n such that

$$t'(n-1,\alpha,\lambda_0) - t'(n-1,1-\beta, \lambda_1) \geq 0 . \qquad (2)$$

The R function **fnlnf**(fstar,pow,p,gam) in the Appendix is used to find n.

3. ANALYSIS OF DATA WITH NON-DETECTS

In situations where an exposure measurement may be less than a detection limit exact methods have not been developed for the lognormal model. The maximum likelihood principle can be used for parameter estimation, and to obtain large sample equivalents of confidence limits for the mean exposure level, the 100pth percentile, and the exceedance fraction. For a detailed discussion of assumptions, properties, and computational issues related to ML estimation see Cox and Hinkley (1979) and Cohen (1991).

3.1 MAXIMUM LIKLIHOOD ESTIMATION FOR LOGNORMAL DATA WITH NON-DETECTS

For notational convenience the m detected values x_i are listed first followed by the x_i^* indicating non-detects, so that the data are $\mathbf{x} = \{ x_i , i = 1,...,m, x_i^* , i = m + 1,...,n\}$. If x_i^* is the same for each non-detect, this is referred to as a left singly censored sample (Type I censoring) and x^* is the DL. If the x_i^* are different, this is known as randomly (or progressively) left censored data [Cohen (1991) and Schmoyer et al (1996)]. In some situations a value of 0 is recorded when the exposure measurement is less than the DL. In this situation, the value of x_i^* is the DL indicating that x_i is in the interval $(0, x_i^*)$. The probability density function for lognormal distribution is

$$g(x;\mu,\sigma) = \exp[- \tfrac{1}{2}(\log(x) - \mu)^2/\sigma^2] \, (\sqrt{2\pi}\,\sigma x)^{-1}, \qquad (3)$$

where y = log(x) is normally distributed with mean μ and standard deviation σ (Aitchison and Brown, 1969). The geometric mean of X is GM= exp(μ) and the geometric standard deviation is GSD = exp(σ) [Strom and Stansbury(2000)] for a summary of these and other relationships for lognormal parameters. Assuming the data are a random sample from a lognormal distribution, the log of the likelihood function for the unknown parameters μ and σ given the data is

$$L(\mu,\sigma) = \sum_{i=1}^{m} \log [g(x_i ; \mu, \sigma)] + \sum_{i=m+1}^{n} \log [G (x_i^* ; \mu ,\sigma)], \qquad (4)$$

where G(x*;μ,σ) is the lognormal distribution function, i.e. G(x*;μ,σ) is the probability that x is less than or equal to x*. The ML equations are obtained by differentiating the log-likelihood function (4) with respect to the μ and σ and setting the result equal to 0; i.e., $\partial L(\mu, \sigma) /\partial\mu = 0$, $\partial L (\mu, \sigma) /\partial\sigma = 0$. These equations cannot be solved directly so a Newton-Raphson-type iterative algorithm is often used to find a root of the system of equations. This leads to

$$C(\boldsymbol{\theta}^0)\boldsymbol{\delta}^0 = G(\boldsymbol{\theta}^0) , \qquad (5)$$

where $G(\boldsymbol{\theta}) = [\partial L(\boldsymbol{\theta})/\partial\theta_j]$, $\theta_1 = \mu$ and $\theta_2 = \sigma$, and $C(\boldsymbol{\theta}^0)$ is the 2 x 2 information matrix with elements $c_{jk} = \partial^2 L(\boldsymbol{\theta})/\partial\theta_j\partial\theta_k$, j, k = 1, 2, . Each of the elements in C and G is elevated at the value of an initial estimate

$\boldsymbol{\theta}^0 = (\mu^0, \sigma^0)$. This linear system of equations (5) is solved for $\boldsymbol{\delta}^0$, and the new value $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \boldsymbol{\delta}^0$ is obtained. The procedure is repeated until a stable solution $\hat{\boldsymbol{\theta}}$ is reached; i.e., $G(\hat{\boldsymbol{\theta}}) = 0$ and $C(\hat{\boldsymbol{\theta}})$ is negative definite. The large sample covariance matrix of the ML estimate $\hat{\boldsymbol{\theta}}$ is obtained by inverting the information matrix evaluated at $\hat{\boldsymbol{\theta}}$, i.e. $V(\hat{\boldsymbol{\theta}}) = C(\hat{\boldsymbol{\theta}})^{-1}$. The numerical approach used here is based on the R function **optim,** a general-purpose optimization procedure that includes the Nelder-Mead, quasi-Newton, and conjugate-gradient algorithms. If the algorithm converges (as indicated by the convergence code from **optim**) and $\hat{\boldsymbol{\theta}}$ is an interior point in the parameter space, it is the unique global maximum of (4) for the situation considered here. The ML estimates $\hat{\mu}$, $\hat{\sigma}$, and standard errors are obtained using the R function **mlndln** provided in the Appendix. Note that for complete samples m = n and the second term in equation (1) is not present. In this case, the solution of the likelihood equations result in well known estimate $\hat{\mu} = \Sigma\, y_i /n$, $\hat{\sigma} = [\Sigma(y_i - \hat{\mu}_i)^2 /n]^{1/2}$, where $y_i = \log(x_i)$.

3.2 CONFIDENCE LIMITS FOR THE MEAN EXPOSURE LEVEL WITH NON-DETECTS

To test the hypothesis H_0: $\mu_x \geq \mu_x^*$, at the $\alpha = 1-\gamma$ significance level a one-sided upper $100\gamma\%$ confidence limit is needed. The first method considered is to use the censored data equivalent of Cox's direct method; i.e., calculate $\hat{\varphi} = \hat{\mu} + ½\,\hat{\sigma}^2$, $\text{var}(\hat{\varphi}) = \text{var}(\hat{\mu} + ½\hat{\sigma}^2)$ where

$$\text{var}(\hat{\varphi}) = \text{var}(\hat{\mu}) + ¼\,\text{var}(\hat{\sigma}^2) + \text{cov}(\hat{\mu}, \hat{\sigma}^2). \qquad (6)$$

In (6) $\hat{\mu}$ and $\hat{\sigma}^2$ are the ML estimates of μ and σ^2, and the estimated variances and covariance are obtained from

$$V(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \text{var}(\hat{\mu}) & \text{cov}(\hat{\mu}, \hat{\sigma}^2) \\ \text{cov}(\hat{\mu}, \hat{\sigma}^2) & \text{var}(\hat{\sigma}^2) \end{bmatrix}. \qquad (7)$$

The ML estimate of μ_x is $\exp(\hat{\varphi})$, the $100\gamma\%$LCL for μ_x is $\exp[\hat{\varphi} - t\,\text{var}(\hat{\varphi})]$, and the $100\gamma\%$UCL for μ_x is $\exp[\hat{\varphi} + t\,\text{var}(\hat{\varphi})]$, where $t = t(\gamma, m-1)$. The resulting confidence interval (LCL,UCL) has confidence level $100(2\gamma-1)\%$. An equivalent procedure is to estimate $\varphi = \mu + ½\sigma^2$ and its standard error directly, i.e. by solving (5) with $\theta_1 = \mu + ½\sigma^2$ and $\theta_2 = \sigma^2$. The R function **mlndln** provided in the Appendix returns ML estimates of μ, σ, φ, σ^2, and estimates of the standard errors for each of the parameter.

A second method for obtaining an UCL for μ_x is based on the procedure proposed by Lyles and Kupper (1996) for the complete data case. They use the relationship between the statistics $\bar{y} + cs_y$ and the non-central t distribution to obtain an approximate UCL for $\log(\mu_x)$ of $\bar{y} + \hat{c}_u\, s_y$ where,

$$\hat{c}_u = \left[-\hat{\delta}\sqrt{n/(n-1)}\right] / \chi(\alpha, n-1) + t(1-\alpha, n-1)/\sqrt{n}. \qquad (8)$$

In (8), $\chi(\alpha, n-1)$ is the positive square root of the 100α percentile of the chi-square distribution with n-1 degrees of freedom, and $\hat{\delta} = -\sqrt{n}\, s_y / 2$. The quantity $\hat{c}_u$ is an estimate of the upper bound of $c = -t'(n-1, \alpha, \delta)/\sqrt{n}$ where t' is the 100αth percentile of the noncentral t distribution with n-1 degrees of freedom and non centrality parameter $\delta = -\sqrt{n}\sigma/2$. For censored data, an approximate log(UCL) for μ_x is $\hat{\mu} + \hat{c}_u\, \hat{\sigma}$ where in calculating $\hat{c}_u$ n is replaced with m. The log(LCL) is obtained in

a similar way. We speculate that the 100γ% approximate UCL for μ_x, $\exp(\hat{\mu} + \hat{c}_u\ \hat{\sigma})$ should be a conservative upper bound. When there are no non-detects (i.e. m = n) Lyles and Kupper (1996) have shown that this procedure is similar in terms of power and Type I error rate to Land's exact method in most situations they considered. Recall that the exact method depends on $\hat{\mu}$ and s_y^2 being independent and respectively normally and a constant times a chi-square. For left censored data the $\text{cov}(\hat{\mu}, \hat{\sigma}^2)$ (see equation 6) is negative and increases in magnitude as the proportion of non-detects increases. The R function LKcl computes confidence limits for μ_x using this approximate method.

3.3 CONFIDENCE LIMITS FOR THE PTH PERCENTILE WITH NON-DETECTS

The point estimate of $y_p = \log(X_p)$ is $\hat{y}_p = \hat{\mu} + z_p\ \hat{\sigma}$ with variance

$$\text{var}(\hat{y}_p) = \text{var}(\hat{\mu} + z_p \hat{\sigma}) = \text{var}(\hat{\mu}) + z_p^2 \text{var}(\hat{\sigma}) + 2z_p \text{cov}(\hat{\mu}, \hat{\sigma}). \quad (9)$$

The 100γ% LCL and UCL for X_p are

$$LX(p,\gamma) = \exp[\hat{y}_p - t(\gamma,(m-1))\text{var}(y_p)^{1/2}],$$
$$UX(p,\gamma) = \exp[\hat{y}_p + t(\gamma,(m-1))\text{var}(y_p)^{1/2}]. \quad (10)$$

$UX(p,\gamma)$ is the estimated 100p-100γ geometric upper tolerance limit. The ML estimates of $\text{var}(\hat{\mu})$, $\text{var}(\hat{\sigma})$, and $\text{cov}(\hat{\mu}, \hat{\sigma})$ are obtained from the ML variance-covariance matrix using R function **mlndln** provided in the Appendix. The null hypothesis H_0: $X_p \geq L_p$ is rejected at the $\alpha = (1-\gamma)$ significance level if 100γ% UCL for X_p is less than L_p (indicating the exposure profile is acceptable). A second method that can be used to estimate the upper tolerance limit is to treat $\hat{\mu}$ and $\hat{\sigma}$ as if they were obtained from a complete sample of size m and calculate $UX(p,\gamma) = \exp(\hat{\mu} + K\hat{\sigma})$, where K is obtained from the non-central t distribution using m, p, and γ as described in Section 2.2. If there are no non-detects, then m = n and method 2 provides the exact upper tolerance limit (provided the unbiased estimate of σ^2 is used). The R function **lnclxpnd** at the AEOD web site calculates estimates of $LX(p,\gamma)$ and $UX(p, \gamma)$ using both large sample ML approach (method 1) and using approximate K values (method 2). Method 2 is the result of "analogical reasoning" and it appears to be a conservative upper bound for $UX(p, \gamma)$ for lognormal data with non-detects. The K factor in Section 2.2 is obtained using the fact that $\bar{y}$ and s_y^2 are independent statistics calculated from a random sample from a normal distribution Johnson and Welch (1940). The ML estimates of μ and σ^2 from censored samples do not satisfy these assumptions.

3.4 CONFIDENCE LIMITS FOR EXCEEDANCE FRACTION WITH NON-DETECTS

The ML point estimate of F_L is $f = 1-N(v)$ where $v = [\log(L)-\hat{\mu}]/\hat{\sigma}$. The large sample 100γ% LCL for $V = [\log(L) - \mu]/\sigma$ is $LCLv = v - t(\gamma, m-1)\ \text{var}(v)^{1/2}$, where

$$\text{var}(v) = p_1^2\text{var}(\hat{\mu}) + p_2^2\text{var}(\hat{\sigma}) + 2p_1p_2\text{cov}(\hat{\mu}, \hat{\sigma}), \quad (11)$$

with $p_1 = \partial v/\partial\mu = -1/\hat{\sigma}$ and $p_2 = \partial v/\partial\sigma = -[\log(L)-\hat{\mu}]/\hat{\sigma}^2$. The 100γ% UCL for F_L is $Uf(L,\gamma) = 1-N(LCLv)$.

The 100γ% LCL for F_L is LF(L,γ) = 1-N(UCLv), where UCLv = $u + t(\gamma,m-1)var(v)^{1/2}$. The null hypothesis H_0: $F_L = 1-p$ is rejected if the 100γ% UCL for F_L is less thanF_0 , indicating that the exposure profile is acceptable. The large sample ML estimates of the exceedance fraction and 100γ% confidence limits for lognormal data with non-detects are obtained using the R function **efclnd** in the Appendix.

A second method that can be used to estimate the exceedance fraction and 100γ% confidence limits is to treat $\hat{\mu}$ and $\hat{\sigma}$ as if they were obtained from a complete sample of size m and use the complete sample method described in Section 2.3. Method 2 is the result of "analogical reasoning" and it appears to be a conservative upper bound for Uf(L,γ) for lognormal data with non-detects. The R function **efcl2** in the Appendix is used to calculate these limits**.**

3.5 NON-PARAMETRIC METHODS FOR SAMPLES WITH NON-DETECTS

The product limit estimator (PLE) of the cumulative distribution function was first proposed by Kaplan and Meier (1958) for right censored data. Turnbull (1976) provides a more general treatment of non-parametic estimation of the distribution function for arbitrary censoring. For randomly left censored data, the PLE is defined as follows [Schmoyer et al. (1996)]. Let $a_1 < \ldots < a_M$ be the M distinct values at which detects occur, r_j is the number of detects at a_j , and n_j is the sum of non-detects and detects that are less than or equal to a_j. Then the PLE is defined to be 0 for $0 \le x \le a^o$, where a^o is a_1 or the value of the detection limit for the smallest non-detect if it is less than a_1 . For $a^o \le x < a_M$ the PLE is $\hat{F}_j = \prod_j (n_j - r_j)/n_j$, where the product is over all $a_j > x$, and the PLE is 1 for $x \ge a_M$. When there are only detects this reduces to the usual definition of the empirical cumulative distribution function. The R function **plend** at the AEOD web site is used to compute the PLE.

The PLE is used to determine the plotting positions on the horizontal axis for the censored data version of a theoretical quantile-quantile (q-q) plot for the lognormal distribution (see Chambers et al., 1983). Waller and Turnbull (1992) provide a good overview of q-q plots and other graphical methods for censored data. The lognormal q-q plot is obtained by plotting a_j (on log scale) versus $H_j = G^{-1}(\hat{P}_j)$, where G^{-1} is the inverse of the distribution function of the standard normal distribution and $\hat{P}_j = (\hat{F}_j + \hat{F}_{j-1})/2$ is the plotting position for a_j [Meeker and Escobar (1998, Chap 6)]. Helsel and Cohen (1988) consider alternative procedures that can be used for calculating plotting positions for left censored data. In the complete data case without ties $\hat{P}_j = (j- ½)/n$. If the lognormal distribution is a close approximation to the empirical distribution, the points on the plot will fall near a straight line. An objective evaluation of this is obtained by calculating the square of the correlation coefficient associated with the plot; i.e., $R^2 = cor(loga_j, H_j)^2$. In the complete data case this will be a close approximation to the Shapiro-Wilk W statistic that is used as a test for normality. Verril and Johnson (1988) considers the large sample distribution of the correlation statistic for Type I and Type II right censored data. A formal test for normality of randomly left censored data has not been developed.

The mean ($\overline{x}_p$) of the PLE is a censoring-adjusted point estimate of μ_x. An approximate standard error of the PLE mean can be obtained using the method of Kaplan and Meier (1958) and the 100γ% UCL is $\overline{x}_p + t(\alpha, m-1)\, s_p$, where s_p is the Kaplan-Meier standard error of $\overline{x}_p$ adjusted by the factor m/(m-1), where m is the number of detects in the sample. When there is no censoring this reduces to the second

approximate method described by Land (1972). The R function **kmms** at the AEOD web site is used to calculate $\overline{x}_p$, s_p, and the confidence limits.

3.6 NON-PARAMETRIC UPPER TOLERANCE LIMIT AND EXCEEDANCE FRACTION

A non-parametric upper tolerance limit can be obtained using the method described by Somerville (1958). Given a random sample of size n from a continuous distribution, then, with a confidence level of at least γ, 100p percent of the population will be below the k^{th} largest value in the sample. The maximum non-detect must be less than the k^{th} largest value. The value of k for specific values of n, p, and γ can be obtained from published tables or, for any reasonable values of n, p, and γ, by using the R function **nptl** provided in the Appendix. The 100γ% upper tolerance bound is equivalent to an upper 100γ% confidence limit for the 100pth percentile of the population.

When the distribution function for the Xs is not specified a "nonparametric" approach can be used to estimate F_L the proportion of measurements that exceed the limit L. Given a random sample of size n the number y of nonconforming observations (i.e., y = number of $x_i > L$) is described using the binomial distribution. The point estimate of F_L is f= y/n and confidence limits are obtained using the method of Clopper and Pearson (1934) [Hahn and Meeker (1991, chap 6)] and the R documentation for base R function **binom.test**. The R driver function **efclnp** returns the point estimate of F_L and the lower and upper 100γ% confidence limits for F_L.

4. APPLICATIONS

In several situations of practical interest statistical analysis of left censored data from a lognormal distribution are required. The "exact" results for complete samples described in Section 2 have not been developed for censored data. The methods presented here are "large sample" results and follow directly from the properties of ML estimators described in Section 3. Each of the examples will describe the censored data equivalent of the exact methods used with complete samples. The emphasis here is on describing the methods and software. The focus in the examples is two areas of application that are part of the Department of Energy (DOE) Chronic Beryllium Disease Prevention Program. The DOE is concerned with monitoring objects (e.g., equipment, buildings) for beryllium contamination and workers for exposure to beryllium in the workplace. The first example describes the results of a survey to evaluate possible beryllium "contamination" based on surface wipe sampling of a smelter facility used to recycle metal. The second example describes the results of a beryllium worker-monitoring program using 8-hour TWA. In both situations "limit values" have been established to determine if exposure levels are acceptable; i.e., the object is "clean" or the workplace is "safe." In general, the limit value will depend on the strategy that is being used as described in the introduction. In the examples the null hypothesis of interest is that the 95^{th} percentile of exposure distribution does not exceed the specified limit. Hewett (1996) explains that occupational exposure limits (OELs) are generally single shift limits used for day-to-day risk management that will also constrain long-term, working lifetime exposures of each individual worker to protective levels. OELs are based on health or toxicology studies that establish protective mean exposure levels. A work environment that rarely exceeds the OEL will also maintain mean levels well below the OEL. Day-to-day exposure prevention is achieved through investigations to determine cause and corrective actions for exposure measurements above the OEL.

Ninety-five-percent confidence that fewer than 5% of measurements are above a specified limit is a statistical definition of compliance that has come into widespread use to determine the monitoring efforts needed to demonstrate compliance [see chapter 7, Mulhausen and Damiano (1998)]. In this situation the upper tolerance limit (i.e., the 95% UCL for the 95^{th} percentile) and the UCL for the exceedance fraction

are of primary interest (see Section 2.3) to determine if the exposure distribution is acceptable. The exact results for samples from a lognormal (or normal) distribution described in Section 2 and the Appendices of Mulhausen and Damiano(1998) are based on the assumption of complete samples; i.e., no left censored data. The statistical methods and computer software for the analysis of left censored data described in Section 3 can be used to calculate the censored data equivalent of all of the statistics described by Mulhausen and Damiano (1998). Details describing R and the R driver functions used to obtain these results are described in the Appendix and at the AOED website. All of the R functions can be downloaded from the AOED web site.

In the examples, results obtained using R interactively are shown in a `monospaced font like this (where ">" is the R prompt)`. To duplicate these results in the examples read the Appendix and then visit the AOED web site and complete steps 1-4. Note that Exhibit 1 is listed at the end of readall.R at the AOED web site, and the data frame **SESdata** and character string **IpSESdata** will be in the R working directory.

4.1 EXAMPLE 1. SURFACE WIPE SAMPLES FROM ELEVATED SEMELTER SURFACES

The data in Exhibit 1 are 31 surface wipe samples from elevated surfaces of a smelter with beryllium contamination. Exhibit 1 illustrates one method that can be used to enter data into R in the format required for the ML estimation. This would normally be done by using a text editor to create a file (example1.txt). All characters on a line to the right of the # sign are comments. The data is entered into the R working directory using the R function **source**; i.e, if the file is in the directory(folder) where the R session was initiated, **source** ("example1.txt") will input the vectors x, det, and the data.frame SESdata.

Exhibit 1 of Section 4.1

```
#    Illustrates one method that can be used to enter
#    data into R in the format required for ML estimates
#
#       Surface Wipe Samples from Smelter
     IpSESdata <- "SESdata:  Smelter-Elevated Surfaces "
#    IpSESdata is Character string for Use by qqlognB()
x <- (15,15,15,25,25,40,40,40,45,50,50,70,75,95,100,125,125,
     145,145,150,150,165,270,290,345,395,395,420,495,840,1140)
x <- x/1000         # wipe samples micrograms per 100 cm^2
det<- c(0,0,0,rep(1,28) )   # first three values are censored
SESdata<- data.frame(x,det) # R data frame for mlndln()
#
```

ML estimates of μ, σ, $\log(\mu_x)$, and σ^2 are obtained using :

```
> mlndln(SESdata)
              mu     sigma       logE      sig2      -2Log(L) Conver
mle   -2.2907643 1.2760000 -1.4766777 1.6281796 -12.852885390      0
semle  0.2311395 0.1754489  0.3137301 0.4477474  -0.002005525     28
```

The R function **mlndln** is described in the Appendix and is available at the AOED website. The data in Exhibit 1 are shown graphically in Fiqure 1. ML estimates of μ, and σ are shown in title of the plot. To obtain Figure 1, use the following at the R prompt:

```
>qqlognB(SESdata,IpSESdata,L=0.2,unit = "mug/100 cm^2",p=0.95,gam=0.95)
```

If equipment is being evaluated for release to the public or for non beryllium use the DOE has established a release limit for removable beryllium contamination of $L_{95} = 0.2$ $\mu g/100cm^2$**.** The ML estimate of the

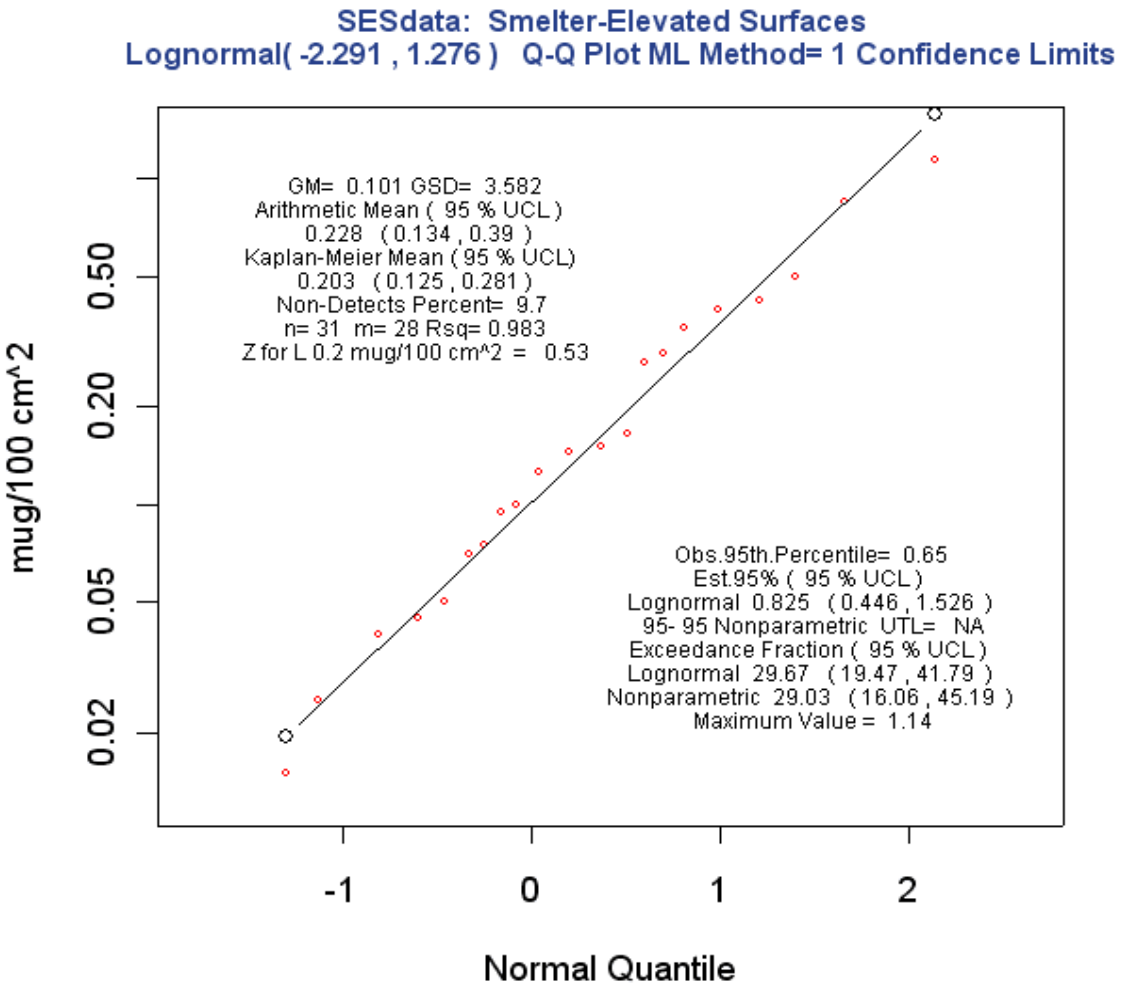

Figure 1. Results for Surface Wipe Samples in Example 1.

exceedance fraction (see Section 2.3), 95% LCL, and 95% UCL are obtained using R function **efclnd** based on the method described in Section 3.4;

```
> efclnd(SESdata,gam=0.95,L=0.2)
   f_MLE LCL_0.95 UCL_0.95
29.66864 19.45963 41.80762
```

The nonparametric estimates of the exceedance fraction, 95% LCL, and 95% UCL are obtained using **efclnp** based on the method described in Section 3.5:

```
> efclnp(SESdata,gam=0.95,L=0.2)
     fnp   LCL_95   UCL_95
29.03226 16.06111 45.19044.
```

The exceedance fraction is an estimate of the percentage of surface area that is expected to exceed the release limit $L_p = 0.2$ μg/100cm^2 with p = 0.95. Both the point estimate and the UCL for F exceed F^o=100(1-p)= 5%, indicating that the equipment is not acceptable. In fact, the 95% LCL indicates that at the 95% confidence level at least 19.5% of the surface area exceeds the release limit. These lognormal based and nonparametric estimates of the exceedance fraction and 95% CLs are shown in the lower right of Figure 1 along with the lognormal based estimate of the 95th percentile $X_p = 0.825$, the lower 95% CL = LX(0.95,0.95) = 0.446, and upper 95% CL= UX(0.95,0.95)= 1.526 (see Section 3.3). The GM, GSD, ML estimates of the (arithmetic) mean of X (with confidence limits) based on the lognormal model, the distribution free Kaplan-Meier mean (with confidence limits), the percent non-detects, the sample size (n), the number of detects (m), R^2 (as defined in Section 3.5), and the z value for the limit L_p are in the upper left corner of Figure 1.

All of the these summary statistics and a brief description of each can be obtained using R function allss(dd,L,p,gam); e.g ,

```
> allss(dd=SESdata,L=0.2,p=0.95,gam=0.95)
           sstat                                              Sec
mu        -2.291 ML estimate of mean of y=log(x)        Sec 3.1
se.mu      0.231 Estimate of standard error of mu       Sec 3.1
   ...
Fnp_0.2   29.030 Nonparmetric estimate of F for limit L Sec 3.6
FnLCL_95  16.060 Nonparmetric estimate of LCL for  F    Sec 3.6
FnUCL_95  45.190 Nonparmetric estimate of UCL for  F    Sec 3.6
>
```

4.2 EXAMPLE 2. TWA BERYLLIUM EXPOSURE DATA

As part of a chronic disease prevention program the DOE adopted an 8-hour TWA OEL limit value of 0.2 micrograms per cubic meter proposed by the American Conference of Government Industrial Hygienists (DOE 10 CRF Part 850 and ACGIH 2004). Figure 2 summarizes the results of 280 personal 8-hour TWA beryllium exposure readings at a DOE facility. This data contains 175 non-detects that range in value from 0.005 to 0.100 μg/m^3. This is an example of random (progressive) left censored data (available at the AOED web site in file beTWA.txt). The q-q plot in Figure 2 is based on the PLE as described in Section 3.5 and was calculated using the R function **plend**(beTWA). Figure 2 can be obtained using R utility function **qqlognB**. To obtain Figure 2, use the following at the R prompt:

```
>beTWA <- read.table("beTWA.txt")
>qqlognB(beTWA,IpbeTWA,L=0.2, unit = "mug/m^3").
```

ML estimates of μ, σ, log(μ_x) , and σ^2 are obtained using :

```
> mlndln(beTWA)
              mu     sigma       logE      sig2      -2Log(L) Conver

     mle   -5.1786787 1.5357165 -3.9994324 2.3585614 -2.175955e+02      0
     semle  0.1340638 0.1155163  0.1485077 0.3548366 -8.918476e-03    105
>
```

The ML estimate of the 95-95 geometric upper tolerance limit is calculate using the results in Section 3.3 equation (9), i.e. $\hat{y}_{.95} = \hat{\mu} + z_{.95}\ \hat{\sigma}$ -2.652 and

$$\begin{aligned} \text{var}(\hat{y}_p) &= \text{Var}(\hat{\mu}) + z_p^2 \text{var}(\hat{\sigma}) + 2z_p \text{cov}(\hat{\mu}, \hat{\sigma}) \\ &= 0.1341^2 + 1.645^2(0.1155)^2 + 2*1.645(-.008918) \\ &= 0.0247 \end{aligned}$$

Then, using equation 6 , UX(0.95,0.95) = exp[-2.652 + 1.659 $(0.0247)^{1/2}$] = 0.091. The nonparametric upper tolerance limit from Section 3.6 is 0.13. Both estimates are well below the limit $L_{0.95} = 0.2$ µg/m^3 indicating that the workplace acceptable (in compliance).

The lognormal based estimate of the exceedance fraction for $L_{0.95} = 0.2$ is 1.01%, and the 95% upper confidence limit Uf(0.2,0.95) = 3.24% < 5% indicating the workplace is acceptable. The non-parametric estimate of the exceedance fraction is 1.43% with 95% UCL = 3.24%. All of the above results provide strong support for the decision that the workplace is in compliance.

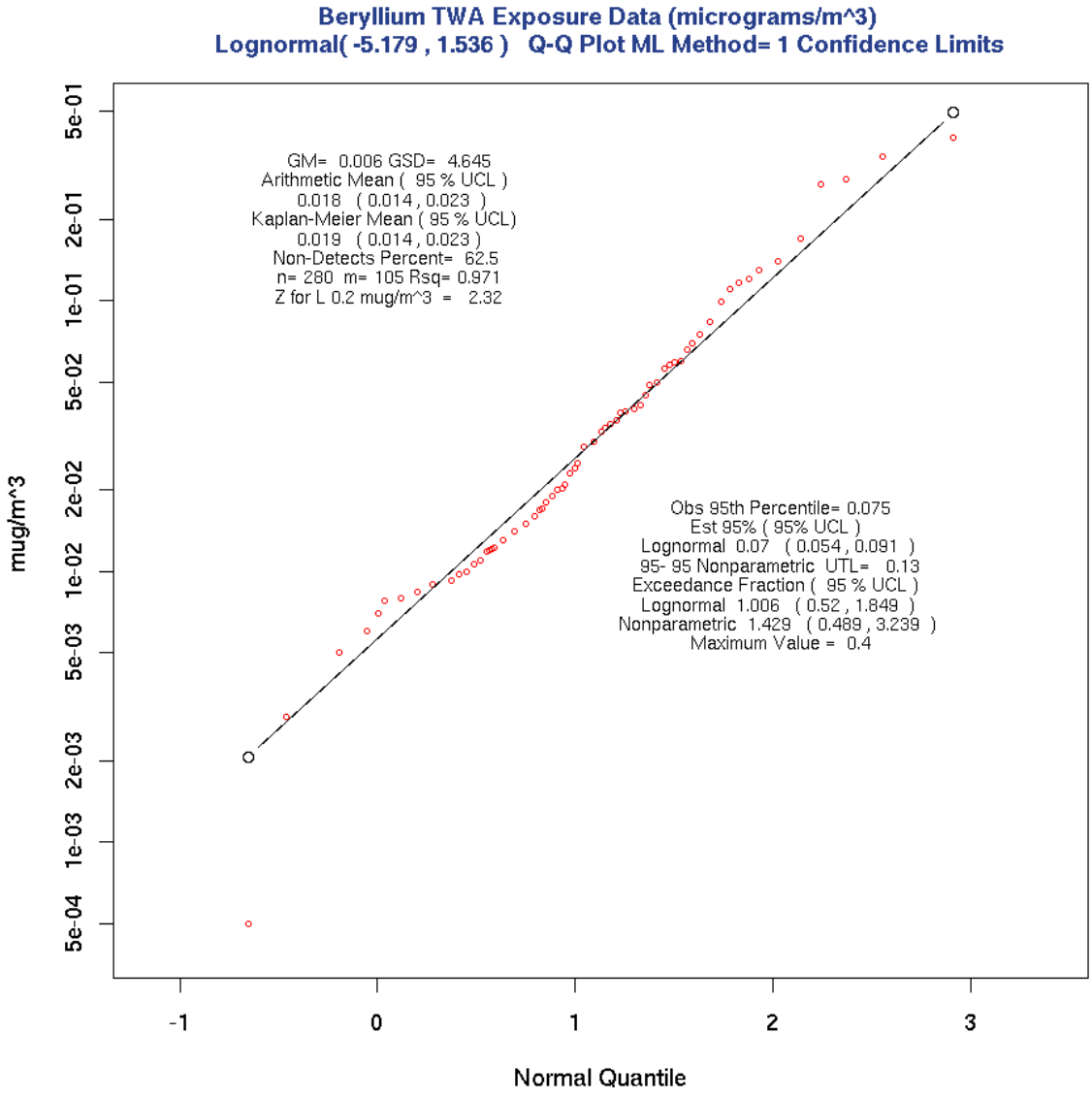

Figure 2. Results for 8-hour TWA Data in Example 2.

5. DISCUSSION

After installing R software as described in the Appendix the R functions posted at the AOED web site can be used to analyze exposure data with non-detects. Data can be entered into R from text files that you create by copying data from word processing, spreadsheet or database files. The output txt files created by R can be imported back into files used for analysis, report writing and record keeping (see the AOED web site, Example 1, and the Appendix for more details). All of the statistics described in the AIHA Consensus Standard are calculated using the methods described in Sections 3 and 4 as illustrated in the examples. For all practical purposes there is no upper limit on the size of the data set that can be analyzed, and as long as there are at least two detected results most of the statistics will be calculated. The R^2 statistic and lognormal q-q plots are equivalent to the W statistic and log probability plots described by Mulhausen and Damiano (1998), and are used in the same way to check on the lognormal assumption and select the appropriate metrics**.**

Note, however, that the results in Sections 3 and 4 are based on large-sample ML methods and the resulting confidence intervals may be "too short" in "small samples." Schmee et al. (1985) have considered confidence limits for the parameters μ and σ for Type I right censored samples from the lognormal distribution. Their report indicates that "exact" results (obtained using Monte Carlo methods) are most useful when the number of uncensored observations is small. They found that when the number of uncensored observations is greater than 20 the agreement between exact and large sample ML confidence limits is good irrespective of the sample size. They did not consider confidence intervals for functions of μ and σ. As far as we know exact (small sample) results have not been developed for randomly (progressively) left censored data.

In this report the UCL for the 95th percentile and the exceedance fraction (both functions of $\hat{\mu}$ and $\hat{\sigma}$) are of primary interest. In the complete data case the upper confidence limits UX(p, γ) and Uf(L_p,γ) always yield equivalent results, i.e. the decision to reject H_0 will always be the same. When the exposure measurements are subject to left censoring the exact methods cannot be used and this is no longer true. Table 1 presents the results from a simulation study that compares the censored ML estimates of UX(p, γ) and Uf(L_p,γ) on the basis of power and type I error rates. The nominal values used were p = 0.95, γ = 0.95, power = 0.8, L_p = 1, and detection limit DL= 0.1. The results in each cell in Table 1 are based on 2000 independently generated random samples from a lognormal distribution under H_0 with F_L= 5%, or separately under H_1 with exceedance fraction F_L = F* and sample size n^* (see columns 2 and 3 in Table I). Columns 3 through 13 in the table are the nominal values for the percent non-detects (PND) in each sample. The sample size was obtained using equation (2) with α = 0.05, (1- β) = 0.8 and F* = 0.75%, 1.5%, and 3%. Figure 3 provides a graphical summary of the relation between sample size and F* for four levels of power. The three vertical line intersect the solid curve for power = 80% at n = 34 for F*= 0.75% , 67 for F* = 1.5%, and 291 for F* = 3%. For each of the 15 combination of values of F* and PND the parameters of the lognormal distribution under H_0 and H_1 and were obtained by solving (with L_p = 1 and DL= 0.1)

$$\mu_0 + z_p \sigma_0 = \log(L_p) \quad ; \quad \mu_0 + z_c \sigma_0 = \log(DL) \ . \tag{11}$$

$$\mu_1 + z_1 \sigma_1 = \log(L_p) \quad ; \quad \mu_1 + z_c \sigma_1 = \log(DL) \ , \tag{12}$$

With z_p = 1.645, z_1 is based on the value of F*, and z_c is determined by the value of PND. For example, with F* = 1.5% and PND = 40% z_1 = 2.17, z_c = -0.2533, μ_1 = -2.062, σ_1 = -0.950, μ_0 = -1.995, and σ_0 = 1.213. This approach requires that for a given value of F* the percent non-detects will be the same under H_0 and H_1. The simulation study was done as follows:

Table 1. Results of Simulation Study with Nominal Value p=0.95

Power = 80%, α = 5%, L_p, and DL = 0.1

			Type I Error[A] Empirical Estimates					Power[B] Empirical Estimates				
			PND[I]					PND				
Procedure	F*[G]	n[H]	1	20	40	60	80	1	20	40	60	80
Uf[C]	0.75	34	6.5	5.2	4.5	3.9	1.1	82	76.1	68.5	55.5	24.9
Uf	1.5	67	5.5	5	5	4.2	4	80.6	76.8	69.1	64.3	50.8
Uf	3	291	5.8	4.5	5.1	5.3	4.8	82.2	79	72.8	67.5	64.8
UX[D]	0.75	34	10.5	8.8	9.8	10.8	11.5	89.1	86.2	82.8	79.1	73
UX	1.5	67	7.6	7	7.6	8.2	10.2	85.8	84.2	78.3	76.5	74.1
UX	3	291	6.3	5.5	6.2	7	8	84.2	81.4	77	73.6	72.8
Method 2[E]	0.75	34	5.4	3.8	2.1	0.8	0.1	79.1	69.8	52.3	30.1	6.8
Method 2	1.5	67	4.8	3.5	2.4	0.7	0.1	78.2	69.5	54.6	35.9	8.9
Method 2	3	291	5.1	3.5	2	1.1	0.2	80.7	73.8	59.8	40.2	11.1
Exact[F]	0.75	34	5.5	4.8	5.2	5.4	4.7	79.8	81.3	79.9	80.5	81.2
Exact	1.5	67	4.9	4.7	4.2	4.8	4.8	79	81	78.6	80	80.9
Exact	3	291	5.1	4.7	4.7	5.4	4.9	81.2	81.8	80.2	79	79.8

[A]The empirical type I error estimate is percent of 2000 tests rejected under H_0
[B]The empirical power estimate is percent of 2000 tests rejected under H_1
[C]The Uf procedure refers to results using Uf(1,0.95) ML method 1
[D] The XU procedure refers to results using UX(0.95,0.95) ML method 1
[E] Method 2 is based on analogical reasoning as describe in the methods section
[F] The exact procedure refers to exact results from the uncensored samples
[G] F* is the value of the exceedance fraction under H_1
[H] n is sample size based on equation 2
[I] PND refers to the nominal percent non-detects

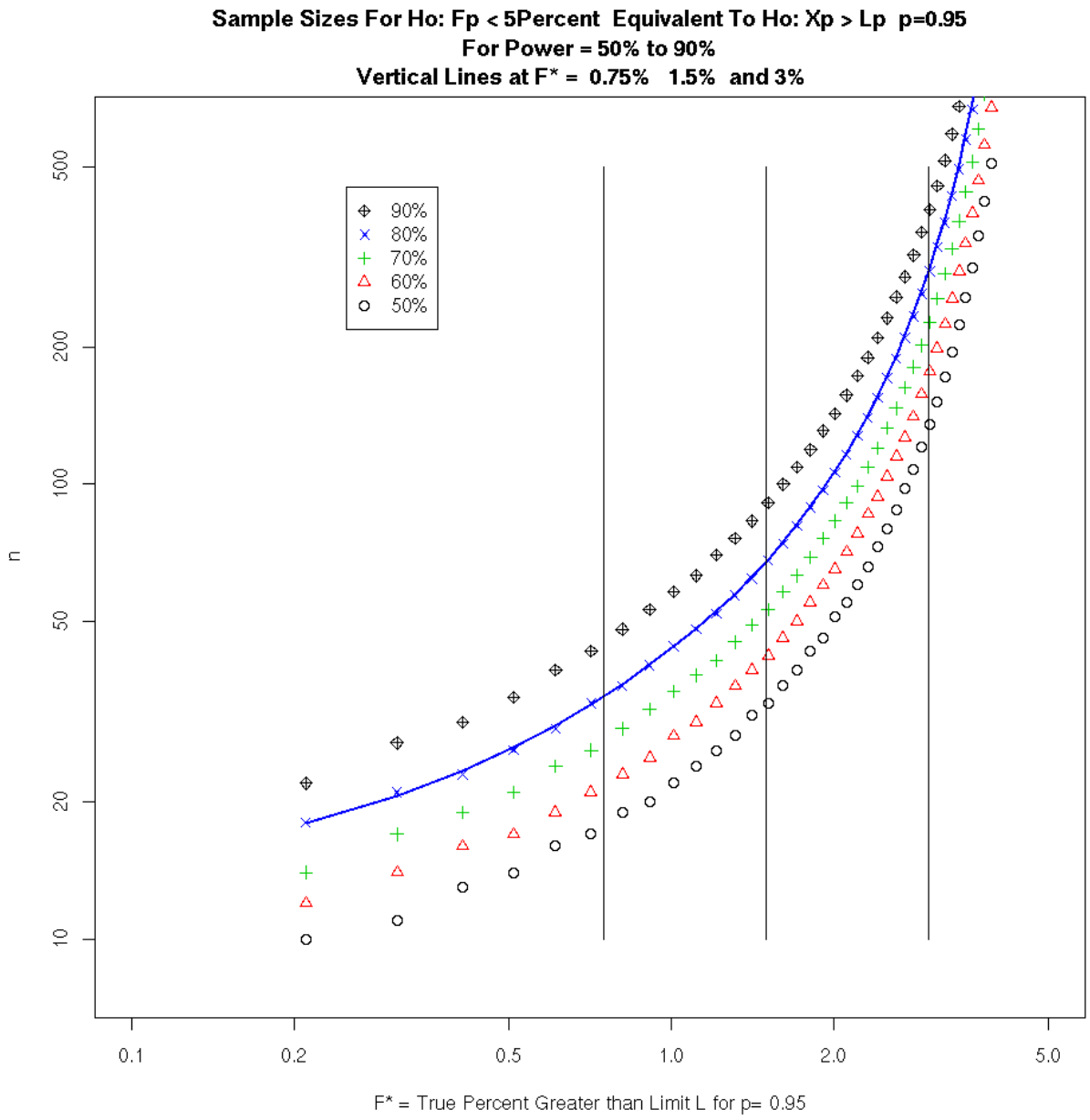

Figure 3. Sample Size Versus F* for Five Levels of Power

For each value of F*, n^* and PND solve equation (11) for μ_0 and σ_0 **;** solve equation (12) for μ_1 and σ_1

(1) Obtain a random sample of size n^* from a lognormal distribution with parameters μ_0 and σ_0. For each complete sample calculate the value of $\bar{y}$ and s_y. Replace each sample value < DL with DL and calculate the ML estimates $\hat{\mu}, \hat{\sigma}$**,** their standard errors, covariance, and m using **mlndln**.

(2) Obtain a random sample of size n^* from a lognormal distribution with parameters μ_1 and σ_1. For each complete sample calculate the value of $\bar{y}$ and s_y. Replace each sample value < DL with DL and calculate the ML estimates $\hat{\mu}, \hat{\sigma}$**,** their standard errors, covariance, and m using **mlndln**.

(3) Save the results from 2 and 3 as a row in a data frame.

(4) Repeat steps 1-4 2000 times.

The results in columns 4 through 8 in Table I are empirical estimates of the Type I error rate (percent of 2000 samples rejecting H_0) for the method in column 1 under H_0 for the sample size in column 3 and PND at the top of the column. The results in columns 9 through 13 are the empirical power estimates, i.e. percent of 2000 tests that rejected H_0 under H_1 with true value $F_L = F^*$ in column 2 and sample size in column 3. The first three rows demonstrate that using the UCL for the exceedance fraction Uf(p,γ) results in a Type I error rate (columns 4 through 8) that is close to or below the nominal level of 5% for censoring levels above 20% . The type I error rate decreases as the PND increase. Columns 9 through 13 of the first three rows show a decrease in power for Uf(p,γ) as a result of increasing levels of censoring. The Type I error rate for upper tolerance limit UX(p, γ) (see Table 1 rows 3-6 columns 4-8 **)** clearly exceeds the nominal level for all situations considered , and consequently the higher levels of power are misleading. The results for method 2 are very conservative; i.e,. the UCLs for the exceedance fraction and 95^{th} percentile provide a conservative upper bound. The last three rows in Table I are based on the exact values of the test statistics, either Uf(p,γ) or UX(p, γ), that were computed from each sample before the sample was censored. These empirical rates are in close agreement with the nominal rates. The average value of the percent non-detect (100*m/n) was calculated for each of the 30 samples and was within 0.1% of the nominal value in every case. These results suggest that in situations where the exceedance fraction is of interest for lognormal data with non-detects the UCL for the exceedance fraction is preferred procedure.

ACKNOWLEDGEMENTS

This work was supported in part by the Office of Environmental Safety and Health, of the U. S. Department of Energy and was performed in the Computer Science and Mathematics Division (CSMD) at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725. Additional funding and oversight have been provided by the National Institute for Occupational Safety and Health (NIOSH) through Contract No ERD-03-2274 from Oak Ridge Associated Universities (ORAU) to support the NIOSH Office of Compensation Analysis and

The authors thank B. Jolene Jones of ORAU's Center for Epidemiologic Research and Lois Thurston in CSMD for their assistance in manuscript preparation. The work has been authored by a contractor of the U.S. Government. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this work, or to allow others to do so for U. S. Government purposes.

REFERENCES

Aitchison, J. and J.A.C. Brown, 1969, The Lognormal Distribution, Cambridge, U.K., Cambridge University Press.

Akritas, M.G., T. F. Ruscitti, and G.S. Patil, 1994. Statistical Analysis of Censored Environmental Data, Handbook of Statistics, Vol. 12, G.P Patil and C.R. Rao (eds), 221-242, Elsevier Science , New York.

American Conference of Governmental Industrial Hygienists (ACGIH), Notice of Intended Change In: 2004 TLVs® and BEIs®, p. 60, ACGIH, Cincinnati, OH.

Armstrong, B. G., 1992. "Confidence Intervals for Arithmetic Means of Lognormally Distributed Exposures," *American Industrial Hygiene Association Journal*, 53(8), 481-485.

Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, 1983. Graphical Methods for Data Analysis, Duxbury Press, Boston.

Clopper, C. J. & Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika_, 26, 404-413.

Cohen, A. C,. 1991. Truncated and Censored Samples. Marcel Dekker, Inc., New York.

Crow, E. L. and K. Shimizu, 1988. Lognormal Distribution, Marcel Decker, New York.

Cox, D. R. and D. V. Hinkley, 1979. Theoretical Statistics. Chapman & Hall, New York.

Cox, D.R. and D. Oakes, 1984. Analysis of Survival Data. Chapman and Hall, New York.

Department of Energy, 10 CFR Part 850, Chronic Beryllium Disease Prevention Program, Federal Register, Vol. 64, No. 235, 68854-68914, December 1999.

Fowlkes, E. B. 1979. "Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions," *Journal of the American Statistical Association*, 74, 561-575.

Hahn, G.J. and W.Q. Meeker, 1991. Statistical Intervals. John Wiley and Sons, New York.

Hewett, P. and G. H. Ganser, 1997. "Simple Procedures for Calculating Confidence Intervals Around the Sample Mean and Exceedance Fraction Derived from Lognormally Distributed Data," *Applied Occupational and Environmental Hygiene*, 12(2), 132-147.

Helsel, D.,1990. "Less Than Obvious: Statistical Treatment of Date Below the Detection Limit", *Environmental Science and Technology,* 24(12), 1767-1774.

Hesel, D.R, and T.A. Cohn ,1988, "Estimation of Descriptive Statistics for Multiply Censored Water Quality Data", Water Resoureces Research, 24, 1997-2004.

Johnson, N. L. and B. L. Welch, 1940. "Application of the Non-Central t-Distribution," *Biometrika*, 31(3/4), 362-389.

Kalbfleisch, J.D. and R.L. Prentice, 1980. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York..

Kaplan, E. L. and P. Meir, P.,1958. "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 457-481.

Land, C. E., 1972. "An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means," *Technometrics*, 14(1), 145-158.

Meeker, W.Q and L.A. Escobar, 1998. Statistical Methods for Reliability Data. John Wiley and Sons, New York.

Moulton, L.H. and N.A. Halsey, 1995. "A Mixture Model with Detection Limits for Regression Analysis of Antibody Response on Vaccine," *Biometrics*, 51, 1570-1578.

Mulhausen, J. R. and J. Damiano, 1998. *A Strategy for Assesing and Managing Occupational Exposures*, Second Edition, AIHA Press, Fairfax, VA.

Neuman, M.C., P.M. Dixon, B.B. Looney,and J.E. Pinder, 1989, "Estimating Mean and Variance for Environmental Samples with Below Detection Limit Observations, Water Resources Bulletin,25, 905-916.

Odeh, R. E. and D. B. Owen, 1980. Tables for Normal Tolerance Limits, *Sampling Plans, and Screening, Marcel* Deker, New York.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Schmee, J., D. Gladstein and W. Nelson, 1985. "Confidence Limits for Parameters of a Normal Distribution From Singly Censored Samples, Using Maximum Likelihood." *Technometrics*, 27, 119-128.

Schmoyer, R. L., J. J. Beauchamp, C. C. Brandt and F. O. Hoffman, Jr., 1996. "Difficulties with the Lognormal Model in Mean Estimation and Testing," *Environmental and Ecological Statistics*, 3, 81-97.

Sommerville, P. N., 1958. "Tables for Obtaining Non-Parametric Confidence Limits," *Annals of Mathematical Statistics*, 29, 599-601.

Taylor, D. J., L. L. Kupper, S. M. Rappaport, and R. H. Lyles, 2001. "A Mixture Model for Occupational Exposure Mean Testing with a Limit of Detection", Biometrics, 57, 681-688.

Tuggle, R. M., 1982. "Assessment of Occupational Exposure Using One-Sided Tolerance Limits," *American Industrial Hygiene Association Journal*, 43, 338-346.

Turnbull, B. W., 1976. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society, Series B (Methodological)*, 38(3), 290-295.

Venables, W. N. and B. D. Ripley, 2002. "*Modern Applied Statistics with S*," 4th edition. Springer-Verlag, New York.

Verrill, S. and R. A. Johnson, 1998. "Tables and Large-Sample Distribution Theory for Censored-Data Correlation Statistics for Testing Normality," *Journal of the American Statistical Association*, 83(404), 1192-1197.

Waller, L. A., and B. W. Turnbull, 1992. "Probability Plotting with Censored Data," *The American Statistician*, 46(1), 5-12.

APPENDIX

R (2004) is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. Detailed documentation on all aspects of R is available at the R home page http://www.r-project.org/. Sources, binaries, and documentation for R can be obtained via CRAN, the "Comprehensive R Archive Network" (click on CRAN on the R home page menu). An Introduction to R and related manuals edited by the R Development Core Team are provided under the "Manuals" link. Additional manuals, tutorials, etc. are provided by users of R under the "Contributed" Link. References are provide under the "Publications" link. [Venables and Ripley(2002)] is a highly recommended book on statistical data analysis using R. All of the R functions discussed in this report and the data used in the examples in Section 4 are available at the website http://www.csm.ornl.gov/esh/aoed (AOED). Most of the serious computing is done by R base functions **optim** and **uniroot**. The R driver functions at AOED are provided to assist the reader that may not have experience with R. They are not "formal" R functions, i.e. there is limited error checking and the usual type of R online "help" files are not provided. Documentation for each function is provided in this report and as comments in the function. All of the files at the AOED web site are ascii (txt) files and can be modified using any text editor (e.g. xemacs, wordpad, vi). The most important functions with more detailed documentation are combined into one file oedmain.R (Exhibit 3). Additional functions that reflect the authors' interest and that may require revisions for other applications are also provided in the file oedutil.R (Exhibit 4). Both of these are available at the AOED web site the AOED website.

In Section 4 several examples of the interactive use of these functions for the analysis of left censored data were provided. These functions can be used for complete data sets by providing an input matrix with the data in column 1 and the censoring indicator for each data value (all equal to 1) in column 2. The results will be the ML estimates. For complete data sets when n is small the "exact " results described in Section 2 may be of interest. The R function **lnexact** in Exhibit 2 illustrates the use of the functions (see Exhibit 3) **extol** and **efcl** for the exact analysis of complete samples. It is used here to provide an introduction to R (see below). These functions are appropriate for complete data sets when the "upper tail" of the lognormal distribution is of interest (see Mulhausen and Damiano, 1998 Appendix VII). In this situation the industrial hygienist picks an upper percentile X_p (often the 95th percentile) and specifies a limit L_p (e.g., the OEL). The value of p is the minimum proportion of the exposure distribution that must fall below the limit L_p for the exposure profile to be considered "acceptable." The exact 100γ% upper confidence limit for the pth percentile X_p is $UX(p, \gamma) = \exp(\bar{y} + K\, s_y)$ and is referred to as the upper tolerance limit (see Section 2.2). The exact 100γ% lower confidence limit for X_p is $LX(p, \gamma) = \exp(\bar{y} + K' s_y)$**.** The point estimate of the exceedance fraction F_L and 100γ% lower and upper confidence limits are calculate using **efcl**(**x**,gam,L) as described in Section 2.3. The Sapiro-Wilk W statistic recommended by AIHA can be used on complete data sets using R function **shapiro.test** for n between 3 and 5000.

To duplicate the results that follow visit the AOED web site and complete steps 1-4. The R working directory AIHD will contain all of the functions described in this report and data frames **aiha** (example data from Mulhausen and Damiano, 1998, page 259) and **aihand** (example data from Mulhausen and Damiano, 1998, page 244 with 3 non-detected values). Note that **aihand** is used as the default data set in all functions that require a two column matrix or data frame (e.g., **mlndln**). The input to **lnexact(x**,p,gam,L) in Exhibit 2 is a vector **x** of positive data x_i, i = 1,…, n , and the values of p, gam(γ), and L. The first line is the function name and arguments. Lines 2 –18 are comments that describe what the function does, the arguments, and the values returned. All characters on a line to the right of the # sign are comments. In a formal R function this information is obtained from a "help" file by entering ? followed by the name of the function; e.g, typing **pt** at the R prompt (the symbol >) describes the probability density function for Student's t distribution that is used by **extol** and **efcl**. Line 20 describes

the error check on line 21. The mean and standard deviation of y = log(x) and sample size are calculated on line 23 (note that the semicolon separates executable statements on the same line). The next 3 lines calculate the point estimate and confidence limits for X_p using **extol** to calculate the values of K and K´. Line 27 combines the three values into a vector **xp** and names it **np**. Line 28 uses **efcl** to calculate the point estimate and confidence limits for the exceedance fraction, and the next line combines vector **xp** and **fp** into a data frame with names based on the input values of p, gam, and L. These functions can be revised using the R functions **fix** or **edit**, or by using a text editor (e.g., Notepad or XEmacs) to make a new file.

Exhibit 2 in the Appendix

```
lnexact <-function(x=aiha[,1],p=0.95,gam=0.95,L=5){
#    Find point estimates and confidence limits for Xp the pth
#    percentile and the exceedance fraction F_L = Pr[ x > L]
#    for a  (complete) sample from a lognormal distribution
# USAGE: lnexact(x,p,gam,L)
# ARGUMENTS:
#     x is a vector of positive lognormal data
#     p is proportion that should fall below L
#     gam is the one-sided confidence level
#     L is the specified limit of interest
# VALUE: data.frame containing point estimates
#     of Xp and F_L in column 1
#     100gam% lower confidence limits in column 2
#     100gam% upper confidence limits in column 3
#      Xp    LX(p,gam)    UX(p,gam)
#      F_L   Lf(L,gam)    Uf(L,gam)
# NOTE: The combined confidence limits are an approximate
#       100*[1 - (1-gam)*2] Percent Confidence Interval
#
#     The next line is an example of "error checking"
 if( any(x) <= 0 ) stop("all data values must be positive")
# calculate mean(yb) and standard deviation(sy) of y=log(x)
yb <- mean(log(x)) ;  sy<- sd(log(x) ) ; n <- length(x)
xp <- exp( yb + qnorm(p)*sy  )        # point estimate of Xp
lxp <- exp( yb + extol(n,p,1-gam)*sy) # exact LCL for Xp
uxp <- exp( yb + extol(n,p,gam)*sy )  # exact UCL for Xp
xp <- c(xp,lxp,uxp)    ; nx <- paste("X",100*p,sep="")
fl <- efcl(x,gam,L,T) ; nf <- paste("F",L,sep="_")
out<-  rbind(xp,fl)
dimnames(out)<- list( c(nx,nf),c("Est","LCL","UCL"))
out
}
```

The following script demonstrates the use of **lnexact**

```
> #   Use data from Hewett and Ganser (HG) 1997 page 135 to
> #   illustrate use of lnexact to obtain exact confidence
> #   limits for Xp percentile and the exceedance fraction
> #   for complete samples from lognormal distribution the
> #   next line assigns the data to the vector variable xhg
>
> xhg<- c(4.25,1.38,3.11,2.20,2.82)
>
> # now use function lnexact with xhg as input
>
> lnexact(xhg,p=0.95,gam=0.95,L=5)
         Est    LCL.95   UCL.95
X95 5.145787 3.6328368 15.10336
F_5 5.744611 0.3795139 35.55304
>
> # HG results--- F_5= 5.7  LCL= 0.38  UCL= 35.55
> #           --- X95  not calculated
>
> # Next use aiha[,1] to check results in Appendix VII
> # of  Mulhausen and Damiano, 1998 (DM)
> Ipaiha
[1] "Strategy for Assessing & Managing Occ Exp page 259"
>
>  lnexact(aiha$x,p=0.95, gam=0.95)
         Est       LCL       UCL
X95 4.842716 3.9015308  7.045908
F_5 4.241070 0.8570198 15.282684
>
>
> #  DM Appendix VII page 278-280 results UTL= 7.11
> #  F_5= 4.4%   95% LCL = 2%   95% UCL  = 15%
> #   difference is due to graphical interpolation and
> #   rounding of the mean and SD of log(x)
> mean( log(aiha[,1]) ) ; sd( log(aiha[,1]) )
[1] 0.9079064 # rounded to 0.91
[1] 0.4070693 # rounded t0 0.41
>
> # DM Appendix VII page 277 mean= 0.91 SD = 0.41
>
```

The results from **lnexact** will agree with those from Odeh and Owen for any reasonable values of p, gam, and L. Note that if the value of L is changed to the 95% UCL for Xp

```
>
> #  change L to 7.046  the 95% UCL for X95
>
> lnexact(aiha[,1],p=0.95, gam=0.95, L=7.046)
              Est     LCL.95   UCL.95
X95     4.8427163 3.90153084 7.045908
F_7.046 0.5143435 0.02923116 4.999718

#  and the 95% UCL for F is 5% as expected
```

This shows the equivalance of the relationship between the upper tolerance limit and the exceedance fraction; i.e, with $\gamma = 0.95$, p= 0.95, L= 7.046, F_0 = 5% the upper tolerance limit is UX(0.95,0.95) = 7.04559 and the upper confidence limit for the exceedance fraction is Uf(7.046,0.95)= 4.99972

H_0: $X_p \geq L$ reject if $UX(p, \gamma) < L$ REJECT H_0
H_0: $F_L \geq 1\text{-}p$ reject if $Uf(L, \gamma) < F_0$ REJECT H_0

Likewise, if the value p = 1 - UCLF_5/100 = .8472 is used

```
> #  change p to  1 - 0.1528= 0.8472
>

> lnexact(aiha[,1],p=0.8472, gam=0.95, L=5)
           Est    LCL.95    UCL.95
X84.72 3.76199 3.1313547  5.000301
F_5    4.24107 0.8570198 15.282684
> #  and the 95% UCL for Xp is 5 as expected
```

This is equivalent to $\gamma = 0.95$, p= 0.8472, L= 5, F_0 = 15.29% , the upper tolerance limit is UX(0.847,0.95)= 4.999157, and the upper confidence limit for the exceedance fraction is Uf(5,0.95) = 15.282684

H_0: $X_p \geq L$ reject if $UX(p, \gamma) < L$ REJECT H_0
H_0: $F_L \geq 1\text{-}p$ reject if $Uf(L, \gamma) < F_0$ REJECT H_0

Exhibit 3 in the Appendix

```
#             oedmain.R Contains R Functions described in
#
#        Statistical Methods and Software for the  Analyis of
#        Occupaional Exposure Data with Non-Detectable Values
#              E. L. Frome and P. F. Wambach
#                 Revision 6a: 29 June 2005
#           http://www.csm.ornl.gov/esh/aoed/
#
#  Name                  Purpose
# -------------------   ------------------------------------------------------
-
#  mlndln(dd)          ML estimates for left censored sample in dd
#  efclnd(dd,gam,L,T) "large sample" CLs for Exceedance Fraction
#  efclnp(dd,gam,L)   non-parametric CLs for F= exceedance fraction
#  extol(n,p,gam)     K factor for exact Lognormal tolerance limit
#  nptl(n,p,gam)      index for Nonparametric tolerance limit
#  efcl(x,gam,L,lx)   exact lognormal CLs for exceedance fraction
#  fnlnf(fs,pow,p,gam) find exact sample size for lognormal
# NOTE:
#  dd  is an  n by 2 matrix with x in column 1 and det(0,1) in column 2
#  gam is confidence level for one sided confidence intervals
#  p   determines the percentile Xp
#  L   is specified limit for Xp
#  fs  true percent of Xs > L
#  pow power of the test
#
# See comments in each function for details
```

```
#
##########################  mlndln ###################################
mlndln <- function(dd = aihand ){
#    ML estimates for lognormal sample with non-detects see Section 3
# USAGE: mlndln( dd )
# ARGUMENT: matrix dd with x[i] in column 1 and det[i] in col 2
#     x[i] is positive lognormal data
#     det[i]=0 for non-detect ; 1 for detect
# NOTE:
#     y= log(x) is normal with mean mu and standard deviation sigma
#     E(X) = exp( mu + 0.5*sig2)= exp(logE) where sig2 = sigma^2
#     m is number of detects and Conver is convergence check
# VALUE: ML estimates of following in 2 by 6 matrix format:
#     mu    sigma      logE     sig2     -2Log(L)     Conver
#   se.mu  se.sigma  se.logE  se.sig2   cov(mu,sig)   m
# REFERENCE: Cohen, A.C (1991) Truncated and Censored Samples
#            Marcel Decker, New York
#  REQUIREs:  ndln() ndln2() loglikelihood functions for optim()
#             see R help file for details on optim() and dlnorm()
   m <- sum(dd[,2])   #  number of non-detects
#  initial estimate of mu and sig (sigma)
      yt <- ifelse(dd[,2]==0,dd[,1]/2,dd[,1] )
      est <- c( mean(log(yt)), sd(log(yt)) )
# ML estimates  mu and sig
est <- optim(est,ndln, method = c("Nelder-Mead"),xd=dd )$par
cont <- list(parscale=abs(est))
opt1 <- optim(est,ndln ,NULL,  method ="L-BFGS-B",lower=c(-Inf,0.0),
         upper=c(Inf,Inf),cont, hessian=T,xd=dd )
conv1 <- opt1$conv          # convergenc check from optim()
mle <- opt1$par             # ML estimate of mu and sig
vcm <- solve(opt1$hessian)
semle <- sqrt(diag( vcm ))  # standard Errors of mu and sig
cov <- vcm[1,2] #  covariace(mu,sig) needed for Tolerance bound
#
# ML estimate of logE  and sig2 (sigmma^2)
#
 est[1] <- mle[1] + 0.5*mle[2]^2
 est[2] <- mle[2]^2
cont <- list(parscale=abs(est))
opt2 <- optim(est,ndln2 ,NULL,  method ="L-BFGS-B",lower=c(-Inf,0.0),
         upper=c(Inf,Inf),cont, hessian=T,xd=dd )
#  next line adds ML estimate of logE  sig2 -2Log(L) and Conver
#  If Conver is not equal to 0 CHECK RESULTs--- see optim() help

mle <- c(mle,opt2$par, 2*opt2$value,opt2$conv+conv1 )
semle <- c(semle,sqrt(diag(solve(opt2$hessian))),vcm[1,2],m)
names(mle) <- c("mu","sigma","logE","sig2","-2Log(L)","Conver")
out <-rbind( mle,semle)
out
}
#
ndln <- function(p=est,xd){
# -log likelihood functions for optim() in  mlndln()
 mu<-p[1]; sig<-p[2]; x<-xd[,1]
xx<-ifelse(xd[,2]==1,dlnorm(x,mu,sig,log=T) , plnorm(x,mu,sig,log.p=T))
  -sum(xx)
}
#
```

```
ndln2 <- function(p=est,xd){
# -log likelihood functions for optim() in  mlndln()
 mu<-p[1] - 0.5*p[2]; sig<-sqrt(p[2]);x<-xd[,1]
xx<-ifelse(xd[,2]==1,dlnorm(x,mu,sig,log=T) , plnorm(x,mu,sig,log.p=T))
  -sum(xx)
}
############################## efclnd  ################################
efclnd<-function(du=aihand,gam=0.95,L=5,dat=T){
#  Calculate ML estimate of exceedance fraction F= Pr[ x > L]
#  and  "large sample" confidence limits
#  for lognormal data with non-detects see Section 3.4
# USAGE: efclnd(du,gam,L,dat)
# ARGUMENTS: if dat=T  du is an n by 2 matrix or data frame
#            if dat=F  du is matrix from mlndln(du)
#            gam is one-sided confidence level(%)
#            L is specifed limit ( e.g OEL)
# VALUE: ML estimate of exceedance fraction and 100gam% CLs
# NOTE: (Lf,Uf) is a 100*[1 - (1-gam/100)*2] Percent Confidence Interval
    if( dat==F ) me <- du
    if( dat==T ) me <- mlndln(du)
#   see mlndln() comments
mu <- me[1,1] ; sig<- me[1,2] ; LL <- log(L)
u<- ( LL - mu)/sig
pd1 <- (-1/sig) ; pd2 <- (mu - LL)/sig^2
# calculate var of u using method of statistical differentials
su <- sqrt( (pd1*me[2,1])^2 + (pd2*me[2,2])^2 + 2*pd1*pd2*me[2,5] )
m <- me[2,6]  # number of non-detects
u1<- u + qt(gam,m-1)*su  ; u2<- u - qt(gam,m-1)*su
f1<- 1-pnorm(u1) ; f2<- 1- pnorm(u2) ; f<- 1- pnorm(u)
out<- 100*c(f,f1,f2)
names(out)<-c("f_MLE",paste("Lf_",gam,sep=""),paste("Uf_",gam,sep="") )
out
}
############################## efclnp ################################

efclnp<-function(dd=aihand,gam=0.95,L=5){
#  For random sample size n from any distribution
#  exceedance fraction F= Pr[ x > L]  See Section 3.4
#  (FL,FU) is 100gam-percent CI for F
# USAGE: efclnp(dd=aiha,gam=0.95,L=5)
# ARGUMENTS: dd = matrix dd with x[i] in column 1 and det[i] in col 2
#           gam= confidence level one sided
#           and L= Limit Fo null value(%)
# VALUE: estimate f of F and exact 100*[ 1 - 2*gam ] percent
#       Confidence Interval for the exceedance fraction F= Pr[ x > L]
# DETAILS: see R function  binom.test()
# ASSUMPTION: all non-detects are less than L
nx<- sum( ifelse(dd[,1] > L,1,0) )
n<-dim(dd)[1]
ef<- nx/n
# f0 <- 1 - p
clev<- 1 - (1 - gam)*2
tmp<- binom.test(nx,n,ef,"two.sided",clev)
fcl<- as.numeric( unlist(tmp[4]) )
out<- 100*c(ef,fcl)
names(out)<-c("fnp",paste("LCL_",100*gam,sep=""),
            paste("UCL_",100*gam,sep="") )
out
}
```

```
############################ extol ###################################

extol<- function(n=50,p=0.95,gamma=0.95) {
# Find K factor for exact tolerance limit for complete sample of
# size n from normal distribution.  See Section 2.2
# USAGE: extol(n,p,gam)
# ARGUMENTS: n: sample size
#            p: quantile of standard normal
#            gamma:  confidence level for one-sided interval
# VALUE: factor K for exact tolerance limit
# DETAILS: R function uniroot is used to find quantile
#     of noncentral t distribution
# NOTE: If M is sample mean and SD is standard deviation
# Prob[ at least 100p% of Xs < M + K*SD] = gamma
# REFERENCES:
#        Johnson, N. L. and Welch, B. L. (1940), Applications of
#        the Non-Central T distribution, Biometrika, 362-389
#
#        Odeh, R.E. and Owen, D.B.(1980) Tables for Normal
#        Tolerance Limits,Sampling Plans, and Screening,
#         Marcel Deker, New York (see Table 1)
#
tx<- function(x,nn=n,th=p,ga=gamma)
{pt(x,nn-1,(-sqrt(nn)*qnorm(th))) + ga- 1}
uout<- uniroot(tx,sqrt(n)*c( -(1/(1- max(p,gamma) )),50) )
K<- -uout$root/sqrt(n)
K
}



############################ nptl ####################################

 nptl<- function(n=100,p=0.95,gamma=0.95){
#  function nptl(n,p,gam)
#     For a random sample of size n calculate largest value
#      of m such that with  confidence level gamma
#      100p percent of population lies  below the
#      mth largest data value in the sample... see Section 3.6
# USAGE: nptl(n,p,gam)
# ARGUMENTS: n: sample size p: defined above
#            gam:  confidence level for one-sided interva
# VALUE: m
# DETAILS: Requries R function qbeata(p,par1,par2)
# REFERENCES:
#   Sommerville, P.N. (1958) Annals Math Stat pp 599-601
k <- ceiling(n*p)
pv <- qbeta(1-gamma,k,n+1-k)
while( pv < p && k < n+1){
k <- k + 1
if( k == n + 1) next
pv<-qbeta(1-gamma,k,n+1-k)
}
if( k <= n) m<- n+1-k else m<- NA
m
}
```

```
############################ efcl ####################################

efcl<- function(x=aiha[,1] ,gam=0.95,L=5,logx=T) {
#  Calculate F = exceedance fraction for limit L (e.g.OEL)
#  For complete random sample size n from lognormal distribution
#  yb is mean  sd is standard deviation of y=log(x) See Section 2.3
#
# USAGE: efcl(x=ex1,gam=0.95,L=5,logx=T)
# ARGUMENTS:x vector of data
#           gam = confidence level
#           L = Limit for exceedance fraction
#           logx if logx=T use log scale
# VALUE: estimate of F ( as percent)
#        exact 100gam% Uf(L,gam) and Lf(L,gam)
#       for the exceedance fraction F= Pr[ x > L]
# NOTE: ( Uf(L,gam),Lf(L,gam) ) is  100*[ 1 - 2*(1-gam) ] percent
#       Confidence Interval for F see Section 2.3
# DETAILS: R function uniroot is used to find noncentrality
#    parameter of noncentral t distribution to calculate CLs
#    for U= (L-mu)/sigma where F= pnorm(U) see JW Eq 16 p 366
# REFERENCES:
#      Johnson, N. L. and Welch, B. L. (1940), Applications
#      of the Non-Central T distribution, Biometrika, 362-389
n <- length(x)
if( logx ) { y<- log(x) ; LL<- log(L) }
      else { y<- x ; LL<-L }
yb <- mean(y)
sd <- sd(y)
out<- efcl2(yb,sd,n,gam,LL)
names(out) <- c("F_L",paste("Lf(",L,",",100*gam,")",sep=""),
                  paste("Uf(",L,",",100*gam,")",sep="") )
out
}
efcl2<-function(yb=0.91,sd=0.41,n=15,gam=0.95,LL=log(5))
{
# USAGE: efcl(yb,sd,n,gam,L)
# ARGUMENTS: yb= mean  sd= standard deviation n= sample size
#            gam = confidence  L= Limit
# DETAILS: see efcl()
del<- function(ncp,tv=t0,df=n-1,eps=cv)
{pt(tv,df,ncp) - eps }
u<- (LL - yb)/sd  ; t0<- sqrt(n)*u
cv<- gam
#    use JW eq 30 to estimate delta
dap<- t0 - qnorm(gam)*( 1 + t0^2/(2*(n-1)) )^0.5
u2<- uniroot(del,dap*c(1/2,2) )$root
cv<- 1 - gam
dap<- t0 - qnorm(cv)*( 1 + t0^2/(2*(n-1)) )^0.5
u1<- uniroot(del, dap*c(1/2,2))$root
out<-c(u1,u2)/sqrt(n)  # Johnson and Welch eq 16
out<- 100*c( 1-pnorm(u), 1 - pnorm(out) )
out
}
```

```
############################## fnlnf ###################################

fnlnf<-function(fstar=1,power=0.9,p=0.95,gam=0.95){
#   For random sample from lognormal distribution
#   given Ho: Xp > Lp  at the alp = 1 - gam significance level
#         where Xp is 100p percentile and Lp is specified limit
#         Ho: Fp > 100*(1-p)
#         where Fp is the percent of Xs > Lp
#   Reject Ho if Uf(Lp,gam) < Lp OR UX(p,gam)< Lp
#   find exact sample size n to provide power of at least (1-beta)
#   when the true value F is F* (fstar)
#
# USAGE: fnlnf(fstar,power,p,gam)
# ARGUMENTS:
#   fstar is true percent of Xs > Lp
#   power = power of test = (1 -beta)
#   p specifies 100pth percentile of X distribution
#   gam desired confidence level = 1 - alp
# VALUE: n sample size
# NOTE: Based on non-central t distribution see
#     Lyles and Kupper (1996) JAIHA vol 57 6-15 Equation 5
if( fstar > 100*(1-p) )
stop(paste("fstar must be less than",100*(1-p),"percent") )

tinv<-function(n,nc,p){
# Given n (sample size)
# nc(non-centrality parameter)
# p Pr[ t <= x}  find x
tx<- function(x,nn=n,ncp=nc,pv=p)
{pt(x,nn-1,ncp) - pv }
uout<- uniroot(tx,sqrt(n)*c( -1/(1-max(p,0.99)) ,50) )
K<- uout$root
K
}
fnlf2<-function(n,fs=fstar,pow=power,pv=p,ga=gam){
#
ncp0<- -sqrt(n)*qnorm(pv)
ncp1<- -sqrt(n)*qnorm( 1 - fs/100 )
t0<- tinv(n,ncp0,1-ga)
t1<- tinv(n,ncp1,pow)
val<- (t0 - t1)
}
out<- floor(uniroot(fnlf2,c(3,20000))$root+1  )
out
}

############################## end oedmain.R ##############################
```

Exhibit 4 in the Appendix

```
#            oedutil.R   Contains R Functions described in
#
#        Statistical Methods and Software for the Analysis of
#        Occupational Exposure Data with Non-Detectable Values
#              E. L. Frome and P. F. Wambach
#                 Revision 6: 13 May 2005
#            http://www.csm.ornl.gov/esh/aoed/
#
#  Name                  Purpose
# -------------------   ------------------------------------------------------
-
#  plend(dd)              Product Limit Estimate for data with non-detects
#  plquan(ple,qq)         calculate the qth quantile from PLE of CDF
#  lnclxpnd(met,p,gam,T) calculate CLs for lognormal percentile
#  kmms(dd,gam)           calculate Kaplan-Meier mean and CLs
#  kmdif(d1,d2,gam,Ls)    two sample t test with non-detects
#  coxcl(dd,gam)          calculate modified Cox Interval for lognormal mean
#  LKcl(dd,gam)           calculate CLs for lognormal mean: Lyles and Kupper
#  allss(dd,L,p,gam)      calculate summary statistic for left censored sample
#  qqqlogn(dd,Ip)         quick q-q plot for left censored lognormal data
#  qqlognB()              q-q plot for left censored lognormal with statistics
#  readss(fn,L,comma)     read data in spread sheet format
#  helpfn()               list all functions

############################## plend ###################################

plend<-function(dd=aihand){
#    Product Limit Estimate for positive data with  non-detectable
#   ( left censored data) values see Section 3.5
# USAGE: plend(dd)
# ARGUMENTS: dd is an n by 2 matrix or data frame
#       dd[,1]=  exposure variable in column 1
#       dd[,2] = censor ( 0 for non-detect 1 for detect)
# VALUE:   data frame with columns
#          apl    a   ple  n r   surv
#
#      apl[j] is adjusted ple used in q-q plot
#      a(j)   is the  value of jth detect (ordered)
#      ple(j) is product limit estimate of cdf (Kaplin-Meier)
#      n(j) = number of detects or non-detects <= x(j)
#      r(j)=  number of detects at x(j)
#      suv(j) = 1 - ple(j)
# REFERENCES: Schmoyer et al (1996) Environmental and Ecological
#             Statistics,3 81-79
# NOTE: see R package dblcens
nn<- length(dd[,1]); n<-rev( 1: nn )
if( sum(dd[,2]) < 2) stop("At least 2 detects required for PLE")
t1<- as.matrix( cbind( dd[ rev(order(dd[,1]) ),1:2],n ) )
dimnames(t1)<- list(NULL,c("x","cen","n") )
t2<- t1[ t1[,2]==1,]  # select detects
ung<- c(-1, diff(t2[,1] ))
d<- rev( table( t2[,1] ) )
#  columns of t2 are x[j] n[j] d[j] for cen=1
#   IN REVERSE order
t2<- as.matrix( cbind( t2[ ung< 0,],d )[,c(1,3:4)] )
#  x contains detects and nx is number of detects
```

```
x<- dd[ dd[,2]==1,1] ; nx<-length(x)
ndt<- max(x)
# add row if min(x) is a non-detect
if( nx < dim(dd)[1] ){
ndt<- dd[ dd[,2]==0, 1] #  nondetects
if( min(ndt) < min(x) ){
nl<- sum( ifelse(dd[,1] <= min(ndt),1,0) )
 frow<- c(min(ndt),nl,0)
 t2<- rbind(t2,frow)
}}
# ple = Prod( [n[j] - d[j] ]/n[j] )
n1<- dim(t2)[1]
ple0<- cumprod( (t2[,2]-t2[,3])/t2[,2] )
ple<- c(1,ple0[1:n1-1])
apl<- as.matrix( rev( (ple0 + ple)/2 ) )
#dimnames(apl)<-list( NULL,c("apl","a","ple","n", "r") )
suv<- 1 - ple ;  suv<-c(suv[2:n1],1 )
t2<- cbind(t2,ple,suv)
t2<- t2[order(t2[,1]),]
#t2<- data.frame(apl,t2[,c(1,4,2,3,5)])
t2<- cbind(apl,t2[,c(1,4,2,3,5)])
dimnames(t2)<-list( NULL,c("apl","a","ple","n", "r","surv") )
t2<-data.frame(t2)
t2
}
############################## plquan ####################################

plquan<-function(pe=ple,qq=0.95){
# Find the qth quantile from PLE of CDF
# USAGE: plend(pe=ple,qq=0.95)
# ARGUMENTS: pe is data.frame from plend
#            qq is such that qq*100 is percentile
# VALUE: Xq the 100qqth percentile
x <- c(0,pe[,2])
p <- c(0,pe$ple)
j <- 0
for( jj in 1:length(x) ) {j<- j+1
   if( p[jj] > qq)  break}
 xq<- (x[j] - x[j-1])/(p[j] -p[j-1])
 xq<- x[j-1] + xq*( qq - p[j-1])
xq
}

############################## lnclxpnd ####################################

lnclxpnd<-function(ae=aihand,p=0.95,gam=0.95,dat=T){
#  Calculate with confidence level gam that
#  100*p percent of population lies  below the upper bound xpu
#     and above the lower bound xpl so that
#  (xpl,xpu) is an approximate 100*[ 1 - 2*(1-p) ] percent
#  Confidence Interval for the Xp percentile of lognormal distribution
# USAGE: lnclxpnd(ae,p,gam,dat)
# ARGUMENTs: if dat=F  ae is output matrix from mlndln(dd)
#            if dat=T  dd is an n by 2 matrix or data frame
#           p is percentile and
#           gam is confidence level
# VALUE: Xp is pth percentile of lognormal distribution
#        (xpl,xpu) METHOD 1 Confidence Interval for Xp
#        (expl,expu) METHOD 2 Confidence Interval for Xp
```

```
# NOTE: The upper bound xpu is the UTL ( upper tolerance limit)
#   for the pth percentile, i.e UTL-pg
#  METHOD 1 IS LARGE SAMPLE RESULT
#  x is lognormal so y=log(x) is normal(mean=ym,sd=ysd)
#    yp = ym + zp*ysd is pth percentile of y
#    ypu = ym + t(gam,nt1-1 )*zp*sdyq  is upper bound
#  var(yqu) = var(yb) + zp^2*var(ysd) +2*zp*cov(yb,ysd)
#   zp is pth percentile of N(0,1) and t() is quantile of  t distn
# METHOD 2 uses tolerance limit factor fron non-central t distribution
#  these are exact limits when there are no nondetects
# REQUIRESs extol()
if(dat) ae<-mlndln(ae)
mu <- ae[1,1] ; sig <- ae[1,2] ; m <-ae[2,6]
vmu <- ae[2,1]^2 ;  vsig <- ae[2,2]^2
zt <- qnorm(p) ; yp<- mu + zt*sig
#  cov(mu,sig) from mlndln() is in ae[2,5]
sdyp <- sqrt( vmu + zt^2*vsig + 2*zt*ae[2,5] )
xp <- exp(yp)
xpu <- exp( yp + qt(gam,m-1 )*sdyp )
xpl <- exp( yp - qt(gam,m-1 )*sdyp )
#  km <- (log(xpu) -mu)/sig  # km is equivalent to K factor
#  require m > 2
K<-extol( max(3,m),p,gam)[1] # apprx K value
expu <- exp(mu + K*sig)
expl <- exp(mu + sig*extol(m,p,(1 -gam))[1] )
#out<-c(xp,xpl,xpu,expl,expu,km,K,m)
#names(out)<-c("Xp","xpl","xpu","expl","expu","km","K","m")
out<-c(xp,xpl,xpu,expl,expu)
names(out)<-c("Xp","xpl","xpu","expl","expu")
out
}

############################## kmms  ####################################

kmms <- function(dd=aihand,gam=0.95){
#    Kaplan- Meier(KM) estimate of mean and Standard Error of
#    the mean for left censored data  see Section 3.5
# USAGE: kmms(dd,cl)
# ARGUMENTS:  dd[,1]= data(exposure variable)
#          dd[,2] = censor ( 0 for non-detect 1 for detect)
#          gam= one-sided confidence level
# VALUE:
#  KM-mean = KM non-parametric estimate of mean
#  KM-se = standard error of mean (adjusted)
#  KM-L and KM-U lower and upper confidence limits
# NOTE: (kml,kmu) is a 100*[1 - (1-cl)*2] Percent Confidence Interval
# REFERENCES:
#  Kaplan- Meier (1958) J Am Stat Assoc 457-481
#       Turnbull (1976) J Royal Stat Soc 290-295
nn<- length(dd[,1]); n<-rev( 1: nn )
if( sum(dd[,2]) < 2) stop("At least 2 detects required for kmms")
#  t1 columns of x cen n in reverse order
t1<- cbind( dd[ rev(order(dd[,1]) ),1:2],n )
t2<- t1[ t1[,2]==1,]  # select detects
ung<- c(-1, diff(t2[,1] ))
d<- as.vector( rev( table( t2[,1] ) ) )
# t2 keep rows with d GT O
t2<- as.matrix( cbind( t2[ ung< 0,],d )[,c(1,3:4)] )
#  x contains detects and nx is number of detects
```

```
x<- dd[ dd[,2]==1,1] ; nx<-length(x)
ndt<-max(x)
#  if min val of x is non-detect add a row
if( nx < dim(dd)[1] ){
ndt<- dd[ dd[,2]==0, 1] #  nondetects
if( min(ndt) < min(x) ){
nl<- sum( ifelse(dd[,1] <= min(ndt),1,0) )
 frow<-c(min(ndt),nl,0)
 t2<-rbind(t2,frow)
}}
# caculate product limit estimate
  ple<-cumprod( (t2[,2]-t2[,3])/t2[,2] )
if( min(ndt) < min(x) ) ple[ dim(t2)[1] ] <- 0
sv<- 1 - ple
xv<- abs(diff(c(t2[,1],0)) )
a<- sv*xv  # a is area for jth interval
ac<- xv*ple ;n1<-dim(t2)[1]
  # if( t2$d[nl==0 )ac[n1-1]<-0
B<- rev(cumsum(rev(ac)))
t2<- cbind(t2,ple,sv,a,ac,B)
t2<- data.frame( t2[order(t2[,1]),] )
#  compute terms for variance
vb<-ifelse( (t2[,2] -t2[,3])==0 , 0, 1/((t2[,2] -t2[,3])*t2[,2]) )
vb<- t2$d*t2$B^2*vb
t2<-data.frame(t2,vb)
kmm<-sum( t2$a) # Kaplan-Meier mean
kmvb<-sum(vb)   # Kaplan-Meier variance (unadjusted)
kmseu<- sqrt(kmvb); kmse<- kmseu* sqrt(nx/(nx-1) )
#
cd<- qt( gam ,nx-1)*kmse
kml<- kmm - cd ; kmu<- kmm + cd
stat<-c(kmm,kmse,kml,kmu,gam)
names(stat)<-c("KM-mean","KM-se","KM-L","KM-U","CL-one")
stat
}
############################## kmdif  ###################################

kmdif<-function(ds=sua1,dr=ref,gam=0.95,Ls=0.0){
# Compare beryllium concentration is survey unit (SU) with reference area (RA)
#
# Ho is mSU > mRA + Ls   mean in SU concentration NOT acceptable
# Ho    mSU - mRA > Ls
# Ha is mSU < mRA + Ls   mean concentration acceptable
# USAGE: kmdif(ds,dr,gam,Ls)
# ARGUMENTS: ds and dr are n by 2 matricies or data frames
#       ds[,1]=  exposure variable in column 1 for SU
#       ds[,2] = detect ( 0 for non-detect 1 for detect) for Su
#       dr[,1]=  exposure variable in column 1 for RA
#       dr[,2] = detect ( 0 for non-detect 1 for detect) for RA
#       gam is confidence level (one sided) for confidence intervals
#       Ls is value of the "site limit" factor ( default is 0)
#  VALUE:
#      KMdif= KMmean from SU - KMmean from RA
#      sed = standard error of the difference
#      LCL = Lower confidence limit
#      UCL = Upper confidence limit
#      gam = one-sided confidence level(%)
#      stdif = KMdif divided by sed
#      pval = probability observing difference GE KMdif if Ho true
```

```
#      Ls =  value of the "site limit" factor
#   To test Ho calculate the 100*gam% UCL for (mSU - mRA)
#     if (95%) UCL < Ls  Recect Ho (the area is acceptable)
#    use approx method of Satterthwaite  for t test with unequal variances
#  REFERENCE: Snedecor and Cochran 7th Ed. page 97
#  REQUIREs: R function kmms()
kr<- kmms(dr,gam)
ks<- kmms(ds,gam)
kmd<- ks[1] - kr[1] # difference of KM means mSU - mRA
sed<- sqrt( ks[2]^2 + kr[2]^2 )# standard error of difference
# nr is the number of detects in RA ns is number of detects in SU
nr<-sum(dr[,2]) ; ns<- sum(ds[,2])
#
ws<- ks[2]^2  ; wo<-kr[2]^2
# dftp is approximate degrees of freedom for t distribution
dftp<- (ws + wo)^2/( ws^2/(ns-1) + wo^2/(nr-1) )
tp<-  qt(gam,dftp)  # t-value for UCL
stdif<- kmd/sed    #  Standardized difference
# p-value If Ho true  i.e if  mSU - mRA = Ls
pval<- 1 - pt( (kmd - Ls)/sed,dftp ) # p-value

cd<- tp*sed
dl <- kmd - cd ; du <- kmd + cd
out<- c( round(c(kmd,sed,dl,du,gam,stdif,dftp,pval,Ls),5) )
names(out)<-c("KMdif","sed","LL","UL","gam","stdif","df","pval","Ls" )
out
}
############################## coxcl  ####################################

coxcl<-function(dd=aihand,gam=0.95,dat=T){
#    Calculate confidence limits for lognormal mean
#    using equivalent of Cox's direct method modified
#    for non-detects as described in Section 3.2
# USAGE: coxcl(dd,cl,dat)
# ARGUMENTS: if dat=T  dd is an n by 2 matrix or data frame
#            if dat=F  dd is matrix from mlndln(dd)
#            gam is one-sided confidence level
# VALUE: ML estimate of AM= exp(mu +0.5sig^2)
#        AM-L lower confidence limit for AM
#        AM-U upper confidence limit for AM
# NOTE: (AM-L,AM-U) is a 100*[1 - (1-cl)*2] Percent Confidence Interval
    if( dat==F ) ys <- dd
    if( dat==T ) ys<- mlndln(dd)
#   see mlndln() comments ys[1,3] is ML estimate of logED
#    ys[2,3] is standard of logED
m <- ys[2,6]  # m = number of non-detects
lex <- ys[1,3]  # same as mu + 0.5*sig^2
tv <- qt(gam, m-1)
lexl <- lex - tv*ys[2,3]
lexu <- lex + tv*ys[2,3]
out<- c( exp( c(lex,lexl,lexu) ) )
names(out)<-c("AM","AM-L","AM-U")
out
}
############################## LKcl ####################################

LKcl<-function(dd=aihand,gam=0.95,dat=T){
#  Calculate confidence limits for lognormal mean
#  using approximate method described in Section 3.2
```

```
# USAGE: LKcl(dd,cl,dat)
# ARGUMENTS: if dat=T  dd is an n by 2 matrix or data frame
#            if dat=F  du is matrix from mlndln(dd)
#            gam is one-sided confidence level(%)
# VALUE: ML estimate of AM= exp(mu +0.5sig^2)
#        LK-L lower confidence limit for AM
#        LK-U upper confidence limit for AM
# NOTE: (LK-L,LK-U) is a 100*[1 - (1-cl)*2] Percent Confidence Interval
# If there are no non-detects (m=n) then (LK-L,LK-U) will be
# an approximation to Land's exact method
# REFERENCES:
#     Lyles and Kupper (1996) JAIHA vol 57 6-15
ml <- dd
if( dat==T ) ml <- mlndln(dd)
  yb<-ml[1,1] ; sy<-ml[1,2]; vyb<-ml[2,1]^2
#
  n <- ml[2,6]  # number of non-detects
 alp <- 1 - gam
 ch<- sqrt( qchisq(alp,n-1) ) ; clo<-sqrt( qchisq(1-alp,n-1) )
#  see Section 3.2 equation 11
cuh<- (sqrt(n)*(sy/2)*sqrt((n-1)/n))/ch  + qt(1-alp,n-1)/sqrt(n)
cul<- (sqrt(n)*(sy/2)*sqrt((n-1)/n))/clo  + qt(alp,n-1)/sqrt(n)
lex<- yb +0.5*sy^2
lexl<-yb + cul*sy ; kl<-  sqrt(n-1)*(lexl - lex)/sy
lexu<- yb + cuh*sy ; ku<- sqrt(n-1)*(lexu - lex)/sy
out<-c(exp(lex) ,exp(lexl),exp(lexu) )
names(out)<-c("AM","LK-L","LK-U")
out
}

############################## allss  ####################################
allss <- function(dd=aihand,L=5,pc=0.95,gam=0.95,mth=1,ro=3){
#  allss  2 Oct04 Revised from 7 Jul 04
#  calculate all summary statistic for left censored sample x[]
#  from lognormal distribution based on Maximum Likelihood
#  and nonparametric methods
#  INPUT:
#    dd is 2col data frame or matrix  with
#    x  in column 1  and cen=1 for  detect 0 for nondetect column 2
#    gam is confidence level (one sided) for confidence intervals
#    pc is quantile for UTL-pc-gam
#    mth=1 for asymptotic mth=2 "pseudo exact"
# REQUIRES mlndln(dd)  plend() coxcl() lnclxpnd() efclnd() efclnp()
#
nt <- length(dd[,1])
ndet <- sum(dd[,2])
if( ndet < 2) stop("At least 2 detects required for allss")
PCndet <- round( 100*(nt-ndet)/nt,1)
du <- dd[,1:2]

ys <- mlndln(du)
GM <- exp(ys[1,1])
GSD <- exp(ys[1,2])

#  Method 1 is equivalent to modified  Cox direct Method
#  approximate CLs for lognormal mean
if(mth==1) {clo<-coxcl(ys,gam,F)
   EX<-clo[1]
   EXL<-clo[2]; nEXL=paste("LCLa",100*gam,sep="_")
```

```
   EXU<- clo[3]; nEXU=paste("UCLa",100*gam,sep="_")
}

#  If mth=2 Use Modifed Method of Lyles and Kupper
#  to calculate approximate CLs for lognormal mean
if(mth==2) {lko<-LKcl(ys,gam,F)
   EX<-lko[1]
   EXL<-lko[2]; nEXL=paste("LCLe",100*gam,sep="_")
   EXU<- lko[3]; nEXU=paste("UCLe",100*gam,sep="_")
}
#  calculate Product limit estimate of CDF for left censored data
cdf <- plend(du)
#  RH is (progressively left censored version of correlation coef)
#  see Verrill and Johnson JASA 1988 Equation 4.4
#  in no censoring Rsq= RH^2 will be approx equal to Shapiro-Wilk W
Rsq <- cor(qnorm(cdf[,1]),log(cdf[,2]) )^2
#  calculate UTL and CLs method 1 and 2
xtmp<-lnclxpnd(ys,p=pc,gam,F)
pc.obs<- plquan( cdf  , pc)   # find pcth percentile of x
pc.est<-xtmp[1]

pco<-paste("Obs%",100*pc,sep=""); pce<- paste("Est%",100*pc,sep="")
#
if( mth==1 ){TL<-xtmp[2];pTL<-paste("LXpa",100*pc,100*gam,sep="")
TU<-xtmp[3];pTU<-paste("UXpa",100*pc,100*gam,sep="")
 }
#
if( mth==2 ) {TL<-xtmp[4];pTL<-paste("LXpe",100*pc,100*gam,sep="")
TU<-xtmp[5];pTU<-paste("UXpe",100*pc,100*gam,sep="")

 }
zL<- (log(L)-ys[1,1])/ys[1,2]; noel<-paste("z_L_",L,sep="")

xmax<-max(du[,1])
m5 <- nptl(nt,pc,gam)
if( is.na(m5) )NPTL<-NA    else NPTL<- rev(sort(du[,1]))[m5]
#
km<-kmms(dd,gam)
nKML<-paste("KLCL",100*gam,sep="_"); nKMU<-paste("KUCL",100*gam,sep="_")
nNPTL<-paste("NpUTL",100*pc,100*gam,sep="")

out<-c(ys[1,1],ys[2,1],ys[1,2],ys[2,2],GM,GSD,EX,EXL,EXU,
      km[1],km[3],km[4],km[2],pc.obs,pc.est,TL,TU,zL)

#  calculate excedance fraction for L lognormal model
#   efclnd() is approximate large sample
if(mth==1) { ef<- efclnd(ys,gam,L,dat=F)
nef1<-paste("Fax",L,sep="_"); nef2<-paste("FaLCL",100*gam,sep="_")
                              nef3<-paste("FaUCL",100*gam,sep="_")
}
if(mth==2){ ef<- efcl2(ys[1,1],ys[1,2],ys[2,6],gam,log(L))
nef1<-paste("Fax",L,sep="_"); nef2<-paste("FeLCL",100*gam,sep="_")
                              nef3<-paste("FeUCL",100*gam,sep="_")
}
#  non-parametric excedance fraction and CLs for L
npf<- efclnp(du,gam,L )
np1<-paste("Fnp",L,sep="_"); np2<-paste("FnLCL",100*gam,sep="_")
                              np3<-paste("FnUCL",100*gam,sep="_")
```

```
sstat<- round( c(out,NPTL,xmax,round(PCndet,2),round(nt),
         Rsq,ndet,ef,npf ),ro )

#

onames<- c("mu","se.mu","sigma","se.sigma","GM","GSD","EX",nEXL,nEXU,

"KMmean",nKML,nKMU,"KM.se",pco,pce,pTL,pTU,noel,nNPTL,"Maximum","NonDet%",
     "n","Rsq","m",nef1,nef2,nef3,np1,np2,np3 )
Sec<-c(
"ML estimate of mean of y=log(x)        Sec 3.1 ",
"Estimate of standard error of mu       Sec 3.1 ",
"ML estimate of sigma                   Sec 3.1 ",
"Estimate of standard error of sigma    Sec 3.1 ",
"MLE of geometric mean                  Sec 3.1 ",
"MLE of geometric standard deviation    Sec 3.1 ",
"MLE of the EX the (arithmetic) mean    Sec 3.2 ",
"Lower Confidence Limit for EX          Sec 3.2 ",
"Upper Confidence Limit for EX          Sec 3.2 ",
"Kaplan-Meier(KM) Estimate of EX        Sec 3.5 ",
"KM Lower Confidence Limit for EX       Sec 3.5 ",
"KM Upper Confidence Limit for EX       Sec 3.5 ",
"Standard Eror of KMmean                Sec 3.5 ",
"Obsevred Percentile of data            Sec 3.5 ",
"ML estimate of Xp the pth percentile   Sec 3.3 ",
"MLE of LX(p,gam) LCL for Xp            Sec 3.5 ",
"MLE of UX(p,gam) UCL for Xp            Sec 3.5 ",
"MLE of the Z value for limit L         Sec 3.4 ",
"Nonparmetric estimate of the UTL       Sec 3.5 ",
"Largest value in the data set                  ",
"The percent of Xs that are left censored       ",
"The number of observations in the data set     ",
"Square of correlation for the q-q      Sec 3.5 ",
"The number Xs greater than the LOD             ",
"MLE of exceedance fraction F for limit L       ",
"LCf(L,gam) MLE of LCL for F            Sec 3.4 ",
"UCf(L,gam) MLE of UCL for F            Sec 3.4 ",
"Nonparmetric estimate of F for limit L Sec 3.6 ",
"Nonparmetric estimate of LCL for  F    Sec 3.6 ",
"Nonparmetric estimate of UCL for  F    Sec 3.6 ",
)

#tmp<<- data.frame(out[1:14],odes,row.names=onames[1:14])

#names(sstat) <- c("mu","se.mu","sigma","se.sigma","GM","GSD","EX",nEXL,nEXU,
#      "KMmean",nKML,nKMU,pco,pce,pTL,pTU,noel,nNPTL,"Maximum","NonDet%",
#     "n","Rsq","m",nef1,nef2,nef3,np1,np2,np3 )
names(sstat) <- onames
sstat<-data.frame(sstat,Sec,row.names=onames)
#sstat<-cbind(sstat)
sstat
}
############################## qqqlogn  ####################################
qqqlogn<-function(dd=aihand,Ip="NONE"){
#
# Quick q-q plot for Left censored sample from  Lognormal Distribution
#  without detailed summary statistics in the plot
#
# USAGE: qqqlogn(dd,Ip)
```

```
# ARGUMENTs: matrix dd with x[i] in column 1 and det[i] in col 2
#     x[i] is positive lognormal data
#     det[i]=0 for non-detect ; 1 for detect
#     Ip  part of title used in plots ( default="NONE")
# VALUE: ML estimates from mlndln()
# REQUIRES: mlndln() and plend()
#
if( sum(dd[,2]) < 2) stop("At least 2 detects required for qqqlogn")
if(Ip=="NONE") Ip<-paste("Lognomal Q-Q Plot for",deparse(substitute(dd) ) )
yvalue <- names(dd)[1]
ple<- plend(dd)      # calculate ple
ym <- ple[,2]        # data on the vertical axis
xq<- qnorm(ple$apl)  # normal quantiles on horizontal axis
#
plot( xq,ym,type = "n",xlab="Normal Quantile",ylab=yvalue,log="y" )
points(xq,ym, pch = 1, cex = 0.6,col="red")
#
#  add results from MLE
mle<-mlndln(dd)
GM<- exp( mle[1,1] )
GSD<- exp( mle[1,2] )
Rsq <- round(cor(qnorm(ple[,1]),log(ple[,2]) )^2,3)
xx<- c(xq[1],xq[length(xq)]);  yhat<- exp(mle[1,1] + xx*mle[1,2])
lines(cbind(xx,yhat),type="b")
# add a title
title(paste(Ip,"GM= ",round(GM,2)," GSD= ",round(GSD,2),"Rsq= ",Rsq) )
mle
}

############################## qqlognB  ###################################

qqlognB<-function(dd=aihand,Ip="NONE",L=5,unit="mg/m^3",pc=0.95,
                  gam=0.95,mth=1,ro=3,loc=F){
#
#  q-q plot for Left censored sample from  Lognormal Distribution
#  with detailed summary statistics in the plot
# USAGE: qqlognB(dd,Ip,L,unit,pc,gam,mth,ro)
# ARGUMENTs: dd two column matris with nonnegative data in d[,1]
#          censoring(0=non-detect;1=detect) indicator in d[,2]
#       Ip  part of title used in plots ( default="NONE")
#       L is specified limit in data units ( e.g OEL) default 5
#       unit is data units--default micrograms/m^3
#       gam is confidence level (one sided) for confidence intervals
#       pc is quantile for Xp and UTL-pc-gam
#       mth = 1 (default) for asymptotic ML  mth=2 "pseudo-exact"
#       ro controls rounding ( digidt to right of decimal point)
#       loc if true use cursor to locate statistics
#  VALUE: Lognormal q-q plot in graphics window with summary statistics
# REQUIRES:  allss()  AND all functions listed in allss()
if( sum(dd[,2]) < 2) stop("At least 2 detects required for qqlognB")
if(loc) cat("use cursor LEFT CLICK to locate statistics \n\ ")
du <- dd[,1:2]
if(Ip=="NONE") Ip<-paste("Lognomal Q-Q Plot for",deparse(substitute(dd) ) )
par(mfrow = c(1, 1), oma = c(1.1, 1.1, 1.1, 1.1), cex = 1)
kmg<- plend(du)
ym <- kmg[,2] ; cl<- 100*gam
tmp<-  allss(du,L,pc,gam,mth,ro)
ss<-as.list(round(tmp[,1],ro))
names(ss)<- row.names(tmp)
```

```
Rsq<-round(ss$Rsq,3)
xq<- qnorm(kmg$apl)
#  scale factors for axis
sf<- 1.2 ;  scy<-c(min(ym)/sf,max(ym)*sf)
#scx<- c(min(xq)/ifelse( min(xq)<sf, max(xq)*sf)
scx<- c( min(xq) - 0.5, max(xq) + 0.5 )
  plot( xq,ym,type = "n",xlim = scx, ylim= scy,
        xlab="Normal Quantile",ylab=paste(" ",unit),log="y" )
points(xq,ym, pch = 1, cex = 0.6,col="red")
xx<- c(xq[1],xq[length(xq)]);  yhat<- exp(ss$mu + xx*ss$sigma)
lines(cbind(xx,yhat),type="b")

# Use results in ss from allss() to add statistics to the plot

tfile<-paste(
   "\n GM= ",round(ss$GM,ro),"GSD= ",round(ss$GSD,ro),
   "\nArithmetic Mean ( ",cl,"% UCL ) ",
   "\n ", round( ss$EX,ro) ,"  (", round(ss$LCL,ro),
   ",",round( ss$UCL ,ro)," )",
   "\nKaplan-Meier Mean (",cl,"% UCL) ",
   "\n", ss$KMmean,"  (",ss$KLCL,",", ss$KUCL,")",
    "\nNon-Detects Percent= ",ss$NonD ,
      "\n n=", ss$n," m=",ss$m,"Rsq=",Rsq )
tadd<- paste("\nZ for L",L,unit," =  ", round( (log(L) -ss$mu)/ss$sigma,2))
if(L > 0) tfile<-paste(tfile,tadd)
      xp<- min(xq) +  0.2* abs( max(xq) - min(xq))
      yp<- min(ym) + 0.5*( max(ym) - min(ym))
if(loc) {xyp<-locator(1); xp<-xyp[1]; yp<-xyp[2]}

 text(xp,yp,tfile,cex=0.7)
 tfile2<- paste( paste( "\nObs ",100*pc,"th Percentile= ", round(ss$Obs,ro),
      "\nEst ",100*pc,"% ( ",cl,"% UCL ) ",sep=""),
 paste("\nLognormal ", round( ss$Est,ro) ,"  (", round(ss$LXp,ro),
   ",",round( ss$UXp ,ro)," )"),paste(
      "\n 95-",cl,"Nonparametric  UTL=  ", ss$NpUTL,
"\nExceedance Fraction ( ",cl,"% UCL ) ",
"\nLognormal ", round( ss$Fax,ro) ,"  (", round(ss$FaL,ro),
   ",",round( ss$FaU ,ro)," )",

"\nNonparametric ", round( ss$Fnp,ro) ,"  (", round(ss$FnL,ro),
   ",",round( ss$FnU ,ro)," )",
      "\nMaximum Value = ",round(max(du[,1]),ro)
       ) )
         xp<- 0.7*max(xq)
      yp<- min(ym) + 0.025*(max(ym) - min(ym))
#yp <- min(ym) + 0.01*(  max(ym)/min(ym) )
if(loc) {xyp<-locator(1); xp<-xyp[1]; yp<-xyp[2]}

text(xp,yp,tfile2,cex=0.7)
      Ip<-paste(Ip,"\n   Lognormal(",ss$mu,",",ss$sigma,")",
         "  Q-Q Plot ML Method=",mth,"Confidence Limits" )
#     "MLEs Method",mth,"Confidence Limits" )
#   navyblue
      mtext( side = 3, line = 1, Ip , cex = 1.0,col="royalblue4",font=2)
      t1<-paste("qqlognB ",substring(date(),1,10),substring(date(),20,24))

}
```

```
############################## readss.R ##########################

readss<-function(fn="beTWA",L=0.2,comma=F) {
#  Read data from fn.txt or fn.csv and calculate all summary statistics
#   using allss().   Output results allss() from to a txt file.
# USAGE: readss("fn",L=2,comma=F)
# ARGUMENTS: fn in double quotes is file name without extension
#            L is specified limit value
#            comma is F (for .csv file) or T (for .txt file)
# NOTE:
#    If comma is F
#    read fn.txt a space delimited file with three couluns
#    Col 1 is rowname Col 2 is positive data value Col 3 is 0 or 1
#    the first record must have 2 valid R column names, e.g. x det
# or if comma is T
#    fn.csv is a two column comma delimited text file created
#    from Excel using Save as type CSV(comma delimited)
#    x in column 1 and det( 0 or 1 ) in column two
#    the first record must have 2 valid R column names, e.g. x det
# VALUE: data.frame
#        writes output from allss to file "fnout.txt"
# EXAMPLEs readss("beTWA",L=0.2,comma=F)
#          readss("aihand",L=0.2,comma=T)
# REFERENCE: see R help files for read.table and read.csv
#
if(comma){ tmp<- read.csv( paste(fn,"csv",sep=".") ,T )
stats<- allss(tmp,L)
stats<-cbind(stats)
nout<-paste(fn,"out.txt",sep="")
}
else{ tmp<- read.table( paste(fn,"txt",sep=".") ,T )
stats<- allss(tmp,L)
stats<-cbind(stats)
nout<-paste(fn,"out.txt",sep="")
}
sink(nout)
print(stats)
sink()
tmp
}
```

```
helpfn<-function(h=T){if(h)cat("type helpfn[Enter] without ()/n")
# list of functions
# NOTE:
#  dd  is n by 2 matrix with x in column 1 and det(0,1) in column 2
#  gam is confidence level for one sided confidence intervals
#  p   determines the percentile Xp
#  L   is specified limit for Xp
#  fs  true percent of Xs > L
#  met ML estimates output from mlndln()
#  ple Product limit estimates from plend()
#  pow power of the test
#
#  FUNCTIONS USED FOR DATA WITH NON-DETECTS
#  Name                  Purpose
# -------------------   ------------------------------------------------
#  allss(dd,L,p,gam)    calculate summary statistic for left censored sample
#  coxcl(dd,gam)        calculate modified Cox Interval for lognormal mean
#  efclnd(dd,gam,L,T)   "large sample" CLs for Exceedance Fraction
#  efclnp(dd,gam,L)     non-parametric CLs for F= exceedance fraction
#  kmms(dd,gam)         calculate Kaplan-Meier mean and CLs
#  LKcl(dd,gam)         calculate CLs for lognormal mean: Lyles and Kupper
#  lnclxpnd(met,p,gam,T)calculate confidence intervals for lognormal
percentiles
#  mlndln(dd)           ML estimates for left censored sample in dd
#  plend(dd)            Product Limit Estimate of CDF for data with non-
detects
#  plquan(ple,qq)       calculate the qth quantile from PLE of CDF
#  qqqlogn(dd,Ip)       quick q-q plot for left censored lognormal data
#  qqlognB()            q-q plot for left censored lognormal with statistics
#  readss(fn,L,comma)   read data in spread sheet format
#
#  FUNCTIONS USED FOR COMPLETE DATA
#  efcl(x,gam,L,lx)     exact lognormal CLs for exceedance fraction
#  extol(n,p,gam)       K factor for exact Lognormal tolerance limit
#  fnlnf(fs,pow,p,gam)  find exact sample size for lognormal
#  nptl(n,p,gam)        index for Nonparametric tolerance limit

# TYPE THE NAME OF THE FUNCTION WITHOUT () FOR DETAILS
}


########################## end util.R ##########################
```

DISTRIBUTION LIST

Internal

1. E. L. Frome
2. J. A. Nichols
3. T. Zacharia
4. ORNL Central Research Library
5. ORNL Laboratory Records - RC
6-7. ORNL Laboratory Records - OSTI

Electronic Notification

8. C. Strader (Cliff.Strader@eh.doe.gov)
9. W. Tankersley (TankersB@orau.gov)
10. P. Wambach (Paul.Wambach@eh.doe.gov)
11. D. Weitzman (David.Weitzman@eh.doe.gov)