

Statistical Analysis of Data with Non-Detectable Values

August 2004

Prepared by

**E. L. Frome
Computer Science and Mathematics Division**

**J. W. Watkins
Oak Ridge Associated Universities**

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via the U.S. Department of Energy (DOE) Information Bridge:

Web site: <http://www.osti.gov/bridge>

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-605-6000 (1-800-553-6847)
TDD: 703-487-4639
Fax: 703-605-6900
E-mail: info@ntis.fedworld.gov
Web site: <http://www.ntis.gov/support/ordernowabout.htm>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source:

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Telephone: 865-576-8401
Fax: 865-576-5728
E-mail: reports@adonis.osti.gov
Web site: <http://www.osti.gov/contact.html>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

STATISTICAL ANALYSIS OF DATA WITH NON-DETECTABLE VALUES

E. L. Frome
Computer Science and Mathematics Division

J. P. Watkins
Center for Epidemiologic Research
Oak Ridge Associated Universities

Date Published: August 2004

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

TABLE OF CONTENTS

	Page
Abstract.....	ix
1. Introduction.....	1
2. Statistical Analysis for Complete Samples.....	1
2.1. Confidence Limit for the Mean Exposure Level.....	2
2.2. Upper Confidence Limit For Pth Percentile.....	3
2.3. Prediction Density for Future Outcomes.....	3
2.3.1. Prediction Density without Explanatory Variables.....	4
2.3.2. Prediction Density with Explanatory Variable.....	4
3. Analysis of Data with Non-Detects.....	5
3.1. Maximum Likelihood Estimation for Lognormal Data with Non-Detects.....	5
3.2. Upper Confidence Limit for the Mean Exposure Level with Non-Detects.....	6
3.3. Upper Confidence Limit for Pth Percentile with Non-Detects.....	7
3.4. Prediction Density with Non-Detects.....	8
3.5. Non-Parametric Methods for Samples with Non-Detect.....	8
3.5.1. Non-Parametric Upper Tolerance Limit.....	9
4. Applications.....	9
4.1. Example 1. Quarterly Doses for A Radiation Worker.....	10
4.2. Example 2. Beryllium Exposure Data.....	12
4.3. Example 3. Linear Regression with Non-Detects.....	14
4.4. Example 4. Quarterly Gamma Doses for a Group of Radiation Workers.....	19
5. Discussion.....	19
6. Acknowledgments.....	21
References.....	22
Appendix.....	25

ABSTRACT

Environmental exposure measurements are, in general, positive and may be subject to left censoring, i.e. the measured value is less than a “limit of detection.” In occupational monitoring, strategies for assessing workplace exposures typically focus on the mean exposure level or the probability that any measurement exceeds a limit. A basic problem of interest in environmental risk assessment is to determine if the mean concentration of an analyte is less than a prescribed action level. Parametric methods, used to determine acceptable levels of exposure, are often based on a two parameter lognormal distribution. The mean exposure level and/or an upper percentile (e.g. the 95th percentile) are used to characterize exposure levels, and upper confidence limits are needed to describe the uncertainty in these estimates. In certain situations it is of interest to estimate the probability of observing a future (or “missed”) value of a lognormal variable. 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 the 95th 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 functions are available at <http://www.csm.ornl.gov/~frome/sand/> (SAND).

Key words: lognormal, maximum likelihood, left censored, regression, confidence limits, prediction density, tolerance limit, exposure measurements, nonparametric

1. INTRODUCTION

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” – see e.g. 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. If, for example, all patients enter the trial at the same time and are followed until a specified end date, then those individuals still alive 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.

In this report, the dependent or response variable of interest is the amount, say D , of a measured quantity. D 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 less than a “limit of detection” say d_i^* (i.e. it is only known that D_i is less than d_i^*). Schmoyer et al. (1996) considered the lognormal model for contaminant concentrations in environmental risk assessment. Another general area of application of the lognormal model is occupational exposure data. Lyles and Kupper (1996) have discussed strategies for the assessment of workplace exposures using time-weighted 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 i) the null hypothesis that the mean exposure level for a similar exposure group (SEG) is below a certain limit, i.e., the long term average permissible exposure limit, or ii) that a specified percentile of the TWA distribution does not exceed a limit. These and other related procedures are described in detail by Mulhausen and Damiano (1998).

2. STATISTICAL ANALYSIS FOR COMPLETE SAMPLES

To review what is known for the complete data case suppose that d_i , $i = 1, \dots, n$ is a random sample from a lognormal distribution with mean $\mu_d = \exp(\mu_y + \sigma_y^2 / 2)$, where μ_y and σ_y^2 are the corresponding mean and variance of $y_i = \ln(d_i)$. Let $\bar{y} = \sum_i y_i / n$ and $s_y^2 = \sum (y_i - \bar{y})^2 / (n-1)$ where s_y^2 is the unbiased estimator for σ_y^2 . The maximum likelihood estimator of σ_y^2 is

$$\hat{\sigma}_y^2 = s_y^2 [(n-1)/n].$$

2.1. CONFIDENCE LIMIT FOR THE MEAN EXPOSURE LEVEL

To test the hypothesis that μ_d is below a specified limit, say μ_d^* , the null and alternative hypothesis are $H_0 : \mu_d \geq \mu_d^*$ vs. $H_1 : \mu_d < \mu_d^*$. A convenient method for testing H_0 (with type I error rate α) is to construct a one-sided upper $(1 - \alpha)100\%$ confidence limit, and reject H_0 if this limit is less than μ_d^* . A number of methods have been described for calculating an upper confidence limit (UCL) for μ_d - see e.g. Armstrong (1992). For Land's (1972) exact method the $(1-\alpha)100\%$ UCL is

$$\exp [(\bar{y} + \frac{1}{2} s_y^2 + C s_y/\sqrt{(n-1)})],$$

where C depends on s_y , n, and α and requires special tables. This is the “best,” i.e. uniformly most powerful unbiased (UMPU), test for complete samples. The “best estimate” of μ_d in complete samples is the minimum variance unbiased estimate (MVUE)--see Hewett and Ganser (1997) for details. Optimal methods (i.e., MVUE or UMPU) for randomly left censored data have not been developed. Two approximate confidence limits have been described by Land (1972) for the complete data case that can be used for censored data.

The first method is attributed to D.R. Cox and is based on calculating an estimate of $\varphi = \ln(\mu_d) = \mu_y + \sigma_y^2 / 2$. For the complete data case the MVUE of φ is $\tilde{\varphi} = \bar{y} + \frac{1}{2} s_y^2$, and the variance of $\tilde{\varphi}$ is $\text{var}(\tilde{\varphi}) = \text{var}(\bar{y}) + \frac{1}{4} \text{var}(s_y^2) = s_y^2/n + \frac{1}{2} s_y^4 / (n-1)$. The $(1-\alpha)100\%$ UCL for μ_d is $\exp [\tilde{\varphi} + t \text{var}(\tilde{\varphi})^{1/2}]$, where $t = t(1-\alpha, n-1)$ is the $100(1-\alpha)$ percentage point of Student's t distribution on n-1 degrees of freedom – see e.g. Land (1972) and Armstrong (1992). The point estimate of μ_d for this method is $\exp(\tilde{\varphi})$. These estimates can be viewed as “bias adjusted”

maximum likelihood (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 $\text{var}(\hat{\varphi}) = \hat{\sigma}_y^2/n + \hat{\sigma}_y^4/(2n)$. The ML estimate of the (arithmetic) mean of d is $\hat{\mu}_d = \exp(\hat{\varphi})$ and the estimate of the $100(1-\alpha)\%$ UCL is $\exp[\hat{\varphi} + t \text{var}(\hat{\varphi})^{1/2}]$. For the censored data case 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 an UCL for μ_d is to calculate the sample mean \bar{d} as the point estimate of μ_d and the approximate $\text{UCL} = \bar{d} + t(1-\alpha, n-1) s_d/\sqrt{n}$, where $s_d^2 = \sum_i (d_i - \bar{d})^2 / (n-1)$. 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{d} and an UCL for μ_d .

2.2. UPPER CONFIDENCE LIMIT FOR Pth PERCENTILE

Let D_p denote the 100pth percentile of the lognormal distribution. The point estimate is $d_p = \exp(\bar{y} + z_p s_y)$ where z_p is the pth quantile of the standard normal distribution. An exact 100 γ % upper confidence limit for the pth percentile is $\hat{U}(p, \gamma) = \exp(\bar{y} + K s_y)$ 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 $-\sqrt{n} z_p$ – see Lyles and Kupper (1996); or Johnson and Welch (1940). The null hypothesis of interest is $H_0: D_p \geq L$ where L is a specified limit (i.e. the occupational exposure limit). If $\hat{U}(p, \gamma) < L$ then reject H_0 indicating the workplace is safe, i.e., the probability is γ (we are 100 γ % confident) that at least 100p% of the d values are less than $U(p, \gamma)$. The R function **extol(n, p, γ)** will return the one-sided tolerance factor K for any reasonable values of n , p , and γ . The function **extol(n, p, (1- γ))** will return the factor K' proposed by Tuggle (1982) that can be used to assess workplace exposure conditions. The one-sided tolerance bounds can be combined to obtain an approximate two-sided tolerance interval which is a confidence interval for D_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 .

2.3. PREDICTION DENSITY FOR FUTURE OUTCOMES

In certain situations, it is of interest to estimate the probability of observing a future (or “missed”) value of a lognormal variable. This situation occurs, for example, when a dose reconstruction is needed for a radiation worker as required under EEOICPA (2000), and there are time periods when an employed worker was not monitored or the dose is “missing.” There are several situations of practical interest that can be considered as special cases of the following general regression model. Let $y_i = \log(d_i)$ denote the observed values of normally distributed random variables with expected value $E(y_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta})$ and variance σ^2 , where the row vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is the i th set of values of p known explanatory (also referred to as independent or predictor variables), and $\boldsymbol{\beta}$ is a p -dimensional vector of unknown parameters. The regression function $\mu(\mathbf{x}, \boldsymbol{\beta})$ relates the expected value of y to the explanatory variables and the parameters. Then, given $\{y_i, \mathbf{x}_i, i = 1, \dots, n\}$ from “the past” we want to estimate the density function of a “future” value, say z , of the response variable at a known future value of the explanatory variables \mathbf{x}_f . That is $p(z; \mathbf{x}_f, \boldsymbol{\beta}, \sigma^2) = n(\mu(\mathbf{x}_f, \boldsymbol{\beta}), \sigma^2)$ where $n(\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 . The “pseudo” prediction density (Geisser, 1971) for z is

$$q(z; \mathbf{x}_f, \mathbf{y}, \mathbf{X}) = n(\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}}), \hat{\sigma}^2), \quad (1)$$

where \mathbf{X} is the known $n \times p$ matrix of explanatory variables, $\mathbf{y} = (y_1, \dots, y_n)$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the ML estimates of $\boldsymbol{\beta}$ and σ^2 , respectively. This pseudo prediction density does not reflect the uncertainty in parameter estimates. A “large sample” maximum likelihood prediction density (MLPD) for $z = \log(d)$, as proposed by Lejeune and Faulkenberry (1982) is

$$q(z; \mathbf{x}_f, \mathbf{y}, \mathbf{X}) = n[\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}}), \hat{\sigma}^2 + \text{var}[\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}})]] , \quad (2)$$

where the second term in the variance of z is the variance of $\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}})$ evaluated at the ML estimate $\hat{\boldsymbol{\beta}}$.

When the mean of y_i is linear in the explanatory variables, i.e., $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y$ and $\hat{\sigma}^2 = (\mathbf{y}-X\hat{\boldsymbol{\beta}})'(\mathbf{y}-X\hat{\boldsymbol{\beta}})/n$ are ML estimates of $\boldsymbol{\beta}$ and σ^2 . The MLPD (2) is then

$$n(\mathbf{x}_f\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 A), \quad (3)$$

where $A = [1 + \mathbf{x}_f(X'X)^{-1}\mathbf{x}_f']$. Levy and Perng (1986) have shown that an “optimal” prediction density for the normal linear model is given by Student’s t density

$$\hat{p}(z; \mathbf{x}_f, \mathbf{y}, X) = t(n-p, \mathbf{x}_f\hat{\boldsymbol{\beta}}, n\hat{\sigma}^2 A/(n-p)), \quad (4)$$

with location parameter $\mathbf{x}_f\hat{\boldsymbol{\beta}}$, dispersion $n\hat{\sigma}^2 A/(n-p)$, and $n-p$ degrees of freedom. The prediction density (4) is the optimal member of a reasonable class of prediction densities based on minimizing the Kullback-Leibler divergence. Note that $[n/(n-p)]\hat{\sigma}^2$ is the bias adjusted estimate of $\hat{\sigma}^2$. The prediction density (4) is equivalent to a particular Bayesian prediction density that is obtained by assuming a diffuse prior for $(\boldsymbol{\beta}, \sigma^2)$ (see e.g. Box and Tiao, 1973). It is also clear that when n is “large” and p is small the MLPD (2) will provide a good approximation to (4) with $\text{var}[\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}})] = \text{var}[\mathbf{x}_f\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2 [\mathbf{x}_f(X'X)^{-1}\mathbf{x}_f']$.

2.3.1. Prediction Density without Explanatory Variables

Let $d_i, i = 1 \dots n$ denote the observed values for a random sample of size n from a lognormal distribution. This is equivalent to a regression model for a “future” or “missed” value of $z = \log(d)$ with $\mathbf{x}_i = 1$ for $i = 1, \dots, n$. Since $A = [1+1/n]$ the prediction density for z from (4) is

$$\begin{aligned} \hat{p}(z; \mathbf{y}) &= t(n-1, \hat{\mu}_y, n\hat{\sigma}_y^2(1+1/n)/(n-1)) \\ &= t(n-1, \bar{y}, s_y^2(1+1/n)), \end{aligned} \quad (5)$$

and for large n the prediction density for d is approximately lognormal.

2.3.2. Prediction Density with Explanatory Variables

Suppose that for each value of d_i there is a known value of the vector \mathbf{x}_i of explanatory variables.

For simple linear regression, let $E(y_i) = \alpha + \beta x_i = \mathbf{x}_i\boldsymbol{\beta}$ where $\mathbf{x}_i = (1, x_i)$ and $\boldsymbol{\beta}' = (\alpha, \beta)$. Then the optimal prediction density for a future or missed value of y at $\mathbf{x} = \mathbf{x}_f$ is obtained using (4) with $\mathbf{x} = (1, x_f)$ and $A = (1 + \mathbf{x}(X'X)^{-1}\mathbf{x}')$, i.e.

$$\hat{p}(z; \mathbf{x}_f, \mathbf{y}, X) = t(n-2, \hat{\alpha} + \hat{\beta} x_f, n\hat{\sigma}^2 A/(n-2)) . \quad (6)$$

Note that since $\text{var}(\hat{\boldsymbol{\beta}}) = s^2(X'X)^{-1}$, where $s^2 = \frac{1}{(n-2)} \sum (y_i - \hat{y}_i)^2$ is the biased adjusted estimate of σ^2 , then the dispersion parameter in (6) is $s^2 + \text{var}(\hat{\alpha} + \hat{\beta} x_f)$ where, $\text{var}(\hat{\alpha} + \hat{\beta} x_f) = \text{var}(\hat{\alpha}) + 2x_f \text{cov}(\hat{\alpha}, \hat{\beta}) + x_f^2 \text{var}(\hat{\beta})$. For large n the prediction density (6) for $z = \log(d)$ is well approximated by the MLPD i.e., $n(\hat{\alpha} + \hat{\beta} x_f, \hat{\sigma}^2 A)$ and the prediction density for d is lognormal.

3. ANALYSIS OF DATA WITH NON-DETECTS

In many situations a sample value may be less than a detection limit that depends on the sampling and analytic methods used. Exact methods have not been developed for the lognormal model with non-detects. The maximum likelihood principle is used for parameter estimation, and to obtain large sample equivalents of confidence limits for the mean exposure level, the pth percentile, and the prediction density. 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 LIKELIHOOD ESTIMATION FOR LOGNORMAL DATA WITH NON-DETECTS

For notational convenience, the m detected values d_i are listed first followed by the d_i^* indicating non-detects, so that the data are $\mathbf{d} = \{d_i, i = 1, \dots, m, d_i^*, i = m + 1, \dots, n\}$ and \mathbf{x}_i is the row vector of explanatory variables for each value of i . If d_i^* is the same for each non-detect, this is referred to as a left singly censored sample (Type I) and d^* is the limit of detection (LOD); if the d_i^* are different, this is known as randomly (or progressively) left censored data – see Cohen (1991) and Schmoyer et al (1996). In some situations (see Example 1) a value of 0 is recorded when the measured dose is less than the LOD. In this situation, the value of $d_i^* = \text{LOD}$ indicating that d_i is in the interval $(0, d_i^*)$. When d_i is a radiation dose (see examples 1 and 3), and the recorded doses is 0, this is sometimes referred to as a “missed dose” and should not be confused with an unmonitored “missed dose.” Assuming the data are a random sample from a lognormal distribution, the log of the likelihood function for the unknown parameters $\boldsymbol{\beta}, \sigma$ given the data is

$$L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^m \log [g(d_i; \mu_i, \sigma)] + \sum_{i=m+1}^n \log [G(d_i^*; \mu_i, \sigma)], \quad (7)$$

where $\mu_i = \mu(\mathbf{x}_i, \boldsymbol{\beta})$, $g(d; \mu, \sigma)$ is the probability density function for lognormal distribution, and $G(d^*; \mu, \sigma)$ is the lognormal cumulative distribution function (CDF), i.e. $G(d^*; \mu, \sigma)$ is the probability that d is less than or equal to d^* . The ML equations are obtained by differentiating the log-likelihood function (7) with respect to the $\beta_j, j = 1, \dots, p$ and σ , i.e.

$$\begin{aligned} \partial L(\boldsymbol{\beta}, \sigma) / \partial \beta_j &= 0, \quad j=1, \dots, p, \\ \partial L(\boldsymbol{\beta}, \sigma) / \partial \sigma &= 0. \end{aligned}$$

These equations cannot be solved directly so a Newton-Raphson type iterative algorithm is used to find a root of this system of equations. This leads to

$$C(\boldsymbol{\theta}^\circ) \boldsymbol{\delta}^\circ = G(\boldsymbol{\theta}^\circ), \quad (8)$$

where $G(\boldsymbol{\theta}) = [\partial L(\boldsymbol{\theta}) / \partial \theta_j]$, $\theta_j = \beta_j, (j=1, \dots, p)$, $\theta_{p+1} = \sigma$, and $C(\boldsymbol{\theta}^\circ)$ is the $(p+1) \times (p+1)$ information matrix with elements $c_{jk} = \partial^2 L(\boldsymbol{\theta}) / \partial \theta_j \partial \theta_k, j, k = 1, \dots, p+1$. Each of the elements in C and G is evaluated at the value of an initial estimate $\boldsymbol{\theta}^\circ = (\boldsymbol{\beta}^\circ, \sigma^\circ)$. This linear system of equations (8) is solved for $\boldsymbol{\delta}^\circ$, and the new value $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^\circ + \boldsymbol{\delta}^\circ$ is obtained. The procedure is repeated until a stable solution $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma})$ 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 $\hat{\theta}$ is an interior point in the parameter space, it is the unique global maximum of (7) for situation considered here, i.e. $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i \boldsymbol{\beta}$ and the m by p matrix of row vector \mathbf{x}_i (the predictor variables for the m detected values) is of full column rank. Detailed instructions on how to obtain and use R are provided in the Appendix. The SAND website also contains the data used in the examples and all of the R driver functions discussed in this report. Note that for complete samples $m = n$ and the second term in equation (7) is not present. In this case, the solution of the likelihood equations result in well known estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $\hat{\sigma}^2 = [\sum(y_i - \hat{\mu}_i)^2 / n]^{1/2}$, where $y_i = \log(d_i)$.

3.2. UPPER CONFIDENCE LIMIT FOR THE MEAN EXPOSURE LEVEL WITH NON-DETECTS

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

$$\text{var}(\hat{\phi}) = \text{var}(\hat{\mu}) + \frac{1}{4} \text{var}(\hat{\sigma}^2) + \text{cov}(\hat{\mu}, \hat{\sigma}^2). \quad (9)$$

In (9) $\hat{\mu}$ and $\hat{\sigma}^2$ are the ML estimates of μ and σ^2 , and the estimated variances and covariance are obtained from

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \text{var}(\hat{\mu}) & \text{cov}(\hat{\mu}, \hat{\sigma}^2) \\ \text{cov}(\hat{\mu}, \hat{\sigma}^2) & \text{var}(\hat{\sigma}^2) \end{bmatrix}. \quad (10)$$

The $(1-\alpha)100\%$ UCL for μ_x is $\exp[\hat{\phi} + t \text{var}(\hat{\phi})]$, where $t = t(1-\alpha, m-1)$.

An equivalent procedure is to estimate $\phi = \mu + \frac{1}{2}\sigma^2$ and its standard error directly, i.e. by solving (8) with $\theta_1 = \mu + \frac{1}{2}\sigma^2$ and $\theta_2 = \sigma^2$. The R function **lnmlnd()** provided in the Appendix returns ML estimates of μ , σ , ϕ , σ^2 , and estimates of the standard errors for each of these parameter.

A second method for obtaining an UCL for μ_d is based on the procedure proposed by Lyles and Kupper (1996) for the complete data case. They use the relationship between the statistics

$\bar{y} + c_{s_y}$ and the noncentral t distribution to obtain an approximate UCL for $\log(\mu_d)$ of $\bar{y} + \hat{c}_u s_y$ where,

$$\hat{c}_u = \left[-\hat{\delta} \sqrt{n/(n-1)} \right] / \chi(\alpha, n-1) + t(1-\alpha, n-1) / \sqrt{n}. \quad (11)$$

In (11), $\chi(\alpha, n-1)$ is the positive square root of the 100α percentile of the chi-square distribution with $n-1$ degrees of freedom, and $\hat{\delta} = -\frac{1}{2}\sqrt{n} s_y$. 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 $100(\alpha)$ th percentile of the noncentral t distribution with $n-1$ degrees of freedom and non centrality parameter $\delta = -\sqrt{n}\sigma/2$. For the

censored data case, the approximate $\log(\text{UCL})$ for μ_d is $\hat{\mu} + \hat{c}_u \hat{\sigma}$ where in calculating \hat{c}_u n is

replaced with m . We speculate that the $(1-\alpha)100\%$ approximate UCL for μ_d , $\exp(\hat{\mu} + \hat{c}_u \hat{\sigma})$ should be a conservative upper bound. When the data is complete (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 $\text{cov}(\hat{\mu}, \hat{\sigma}^2)$ (see equation 9) is negative and increases in magnitude as the proportion of non-detects increases. The R function **LKcl()** computes confidence limits for μ_d using this approximate method.

3.3. UPPER CONFIDENCE LIMIT FOR PTH PERCENTILE WITH NON-DETECTS

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

$$\begin{aligned} \text{var}(\hat{y}_p) &= \text{var}(\hat{\mu} + z_p \hat{\sigma}) \\ &= \text{var}(\hat{\mu}) + z_p^2 \text{var}(\hat{\sigma}) + 2z_p \text{cov}(\hat{\mu}, \hat{\sigma}). \end{aligned}$$

The $100\gamma\%$ UCL for D_p , i.e. the estimated $100p-100\gamma$ geometric tolerance limit is

$$\hat{U}(p, \gamma) = \exp[\hat{y}_p + t(\gamma, m-1)\text{var}(y_p)^{1/2}]. \quad (12)$$

The 100% 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 **lnmlnd()** provided in the Appendix.

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 $\hat{U}(p, \gamma) = \exp(\hat{\mu} + K \hat{\sigma})$, 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 (this requires the bias adjusted estimate of σ). The R function **lnclxpdn()** at the SAND web site calculates estimates of the $U(p, \gamma)$ using both large sample ML approach (method 1) and using K (method 2). Method 2 is the result of analogical reasoning and we view it as a conservative upper bound on $U(p, \gamma)$ 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).

3.4. PREDICTION DENSITY WITH NON-DETECTS

To estimate the prediction density for $z = \log(d)$ at known values of the explanatory variables \mathbf{x}_f , we use the “large sample” MLPD in equation (2), the ML estimate $\hat{\boldsymbol{\theta}}$, and the estimated variance - covariance $V(\hat{\boldsymbol{\theta}})$. If the mean is linear in \mathbf{x} then $\mu(\mathbf{x}_f, \hat{\boldsymbol{\beta}}) = \mathbf{x}_f \hat{\boldsymbol{\beta}}$, and the $\text{var}(\mathbf{x}_f \hat{\boldsymbol{\beta}}) = \mathbf{x}_f V(\hat{\boldsymbol{\beta}}) \mathbf{x}_f'$, where $V(\hat{\boldsymbol{\beta}})$ corresponds to the $p \times p$ submatrix of $V(\hat{\boldsymbol{\theta}})$ obtained by deleting the last row and column. It then follows from large sample results for ML estimators that the prediction density for $z = \log(d)$ is approximately

$$q(z|\mathbf{x}_f) = n(\mathbf{x}_f \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 + \mathbf{x}_f V(\hat{\boldsymbol{\beta}}) \mathbf{x}_f'), \quad (13)$$

i.e. the prediction density for d is lognormal. In particular, if $p=2$, $\boldsymbol{\beta} = (\alpha, \beta)$, and $\mathbf{x} = (1, x_f)$, then

$$\hat{\mu}(\mathbf{x} \hat{\boldsymbol{\beta}}) = \hat{\alpha} + x_f \hat{\beta} \quad \text{and} \quad \text{var}[\hat{\mu}(\mathbf{x} \hat{\boldsymbol{\beta}})] = \text{var}[\hat{\alpha} + x_f \hat{\beta}],$$

$$\text{var}[\hat{\alpha} + x_f \hat{\beta}] = [1, x_f] \begin{bmatrix} \text{var}(\hat{\alpha}) & \text{cov}(\hat{\alpha}, \hat{\beta}) \\ \text{cov}(\hat{\alpha}, \hat{\beta}) & \text{var}(\hat{\beta}) \end{bmatrix} \begin{bmatrix} 1 \\ x_f \end{bmatrix}$$

$$= \text{var}(\hat{\alpha}) + 2 x_f \text{cov}(\hat{\alpha}, \hat{\beta}) + x_f^2 \text{var}(\hat{\beta}),$$

and the MLPD is $n(\hat{\alpha} + \hat{\beta} x_f, \hat{\sigma}^2 + \text{var}[\hat{\alpha} + x_f \hat{\beta}])$.

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-parametric estimation of the distribution function for arbitrary censoring. For randomly left censored data, the PLE is defined as follows – see Schmoeyer et al (1996). Let $a_1 < \dots < a_L$ be the L 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 or detects that are less than or equal to a_j . Then the PLE is defined to be 0 for $0 \leq d \leq a_1'$ where a_1' is a_1 or the value of the detection limit for the smallest non-detect if it is less than a_1 . For $a_1' \leq d < a_L$ the PLE is $\hat{F}_j = \prod_j (n_j - r_j)/n_j$, where the product is over all $a_j > d$, and the PLE is 1 for $d \geq a_L$. Note that when there are only detects this reduces to the usual definition of the cumulative distribution function. The R function **plend()** in the Appendix 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 CDF of the standard normal distribution and $\hat{P}_j = (\hat{F}_j + \hat{F}_{j-1})/2$. 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. $R^2 = \text{cor}(\log a_j, H_j)^2$. 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 (\bar{d}_p) of the PLE is a censoring-adjusted point estimate of μ_d . An approximate standard error of the PLE mean can be obtained using the method of Kaplan and Meier (1958) and the $(1-\alpha)100\%$ UCL is $\bar{d}_p + t(\alpha, m-1) s_p$, where s_p is the Kaplan-Meier standard error of \bar{d}_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()** in the Appendix is used to calculate \bar{d}_p , s_p , and confidence limits.

3.5.1. NON-PARAMETRIC UPPER TOLERANCE LIMIT

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 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 $100\gamma\%$ upper tolerance bound is equivalent to an upper $100\gamma\%$ confidence interval for the $100p$ th percentile of the population.

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 three examples will describe the censored data equivalent of one or more of the exact methods used with complete samples. The emphasis here is on describing the methods and software. More substantive issues will be considered in subsequent reports—see Watkins et al (2004) and Frome and Wambach (2004).

4.1. EXAMPLE 1. QUARTERLY DOSES FOR A RADIATION WORKER

This example demonstrates the use of R for all of the methods described in Section 3. These “informal” or driver functions are provided for the readers’ convenience (see the Appendix for details). In the discussion that follows, we assume that the reader has visited the SAND website and completed Steps 1-4. Typing the name of an R function (or any other object) at the console without the parentheses will list the function (object). The data in Table 1 are an individual’s gamma radiation doses of record at the Y-12 plant in Oak Ridge, TN, from 1961 to 1970. Individuals were monitored quarterly and a recorded dose of zero means the dose to the worker was less than the limit of detection (LOD) unless a smaller value is given—see Watkins et al (2004) for details.

Table 1. Quarterly Film Badge Doses ⁺										
Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Q1	9	6	0	0	0	0	0	25	38	34
Q2	112	182	16	0	29	23	0	80	23	23
Q3	31	4	38	33	22	11	2	0	0	14
Q4	69	143	0	0	66	21	10	10	54	34
⁺ mSv*100										

The driver function **mlnd2**(dd) shown in Exhibit 1 illustrates the use of the R function **optim**() to obtain ML estimates of μ and σ for left censored lognormal data. The input to **mlnd2**(dd) is a two column matrix with the nonnegative data values d_i in column 1, and column two contains a censoring indicator that is equal to zero for non-detects and one for detects. Anything to the right of the # character on a line in Exhibit 1 is a comment. The data from Table 1 in the two column matrix format are available at the SAND website in file Ex1.txt.

Results obtained using R interactively for the data in Table 1 are as follows: (The symbol “>” indicates the R prompt and results obtained interactively are in the font Courier New.)

```
> ex1 <- read.table("Ex1.txt")
> mlnd2(ex1)
       $\mu$        $\sigma$       se( $\mu$ )      se( $\sigma$ )      cov      n      -2Log(L)
3.01279  0.99174  0.17065  0.12883  -0.00407  40.00  280.75718
```

The initial estimates of μ and σ (see Exhibit 1) that are required by **optim**() are the mean and standard deviation (ignoring censoring and dividing non-detects by 2) of $\log(d_i)$. These initial estimates, in the vector **est**, are the first argument to **optim**(). The second argument to **optim**() is the function to be minimized. This function, **nlnd**(), is listed at the bottom of Exhibit 1. The first argument to **nlnd**() is the vector of parameters over which minimization is to take place, and

the second argument is the two column data matrix. **nlnd()** uses base R functions **dlnorm()** and **plnorm()** to compute the values of the lognormal density and cumulative distribution function in the log-likelihood function in equation (1) evaluated at each value of d_i using the current values of μ and σ in **est**. **nlnd()** returns the negative value of the log-likelihood function. The third argument tells **optim()** to use the Nelder-Mead method to find the MLEs of μ and σ , and the last argument contains the data that is passed to **nlnd()**. The Nelder-Mead algorithm is a derivative free robust method that will find the MLE estimates, $\hat{\mu}$ and $\hat{\sigma}$.

Exhibit 1 of Section 4.1

```

mlnd2 <- function(dd =ex1)
{
# mlnd2 find Maximum Likelihood estimates of mu and sig
# left censored sample from lognormal distribution
#
# INPUT: matrix dd with d[i] in column 1 and cen in col 2
# d[i] is positive lognormal data cen=0 for non-detect ; 1 for detect
# y= log(d) is normal with mean mu and standard deviation sig
# OUTPUT: ML estimate of mu and sig estimates of standard errors
# of mu and sig and - 2*Log-likelihood function
#
# REQUIRES: function nlnd() and base R optim()
# initial estimate of mu and sig required by optim()
yt <- ifelse(dd[,2]==0,dd[,1]/2,dd[,1])
est <- c( mean(log(yt)), sd(log(yt) ) )
n <- dim(dd)[[1]] ; m <- sum(dd[,2])
# ML estimate mu and sig
#
est <- optim(est,nlnd, method = c("Nelder-Mead"),dx=dd )$par
cont <- list(parscale=abs(est))
opt1 <- optim(est,nlnd ,NULL, method = "L-BFGS-B",lower=c(-Inf,0.0),
upper=c(Inf,Inf),cont, hessian=T,dx=dd )
mle <- opt1$par # ML estimates of mu and sig
vcm <- solve(opt1$hessian) # ML varaince-covariance matrix
semle <- sqrt(diag(vcm)) # standard errors of mu and sig
cv <- vcm[1,2] # covariace(mu,sig)
est <- c( round( c(mle,semle,cv) ,5 ),n,2*opt1$value)
names(est) <- c("mu", "sig", "se.mu", "se.sig", "cov", "n", "-2Log(L)")
est
}
nlnd <- function(p=est,dx)
{
# compute - log liklihood for lognormal sample
mu <- p[1]; sig <- p[2]; d <- dx[,1]
xx <- ifelse(dx[,2]==1,dlnorm(d,mu,sig,log=T),plnorm(d,mu,sig,log.p=T))
-sum(xx)
}

```

The second call to **optim()** uses a quasi-Newton derivative based method to obtain the Hessian matrix G (the observed information matrix). The inverse of G is the variance-covariance matrix for the ML estimates. The square root of the diagonal terms are estimates of the standard errors of $\hat{\mu}$ and $\hat{\sigma}$, respectively. The off-diagonal element provides an estimate of the covariance of $\hat{\mu}$

and $\hat{\sigma}$. These second order statistics are needed to obtain the ML estimate of the upper tolerance limit and the standard deviation of the prediction density function.

The data in Table 1 are shown graphically in the censored data lognormal q-q plot (see Figure 1) that is obtained using the PLE (see Section 3.4), i.e. columns 1 and 2 from **plend()**:

```
> plend(ex1)
      apl      d      ple      n r      suv
1 0.02109375 2 0.0421875 1 1 1.0000000
2 0.06328125 4 0.0843750 2 1 0.9578125
3 0.10546875 6 0.1265625 3 1 0.9156250
      ....
```

The data points are close to the solid line (which is calculated from the ML estimates), indicating that the lognormal distribution is reasonable model for this data. This is further confirmed by the R^2 of 0.984. The estimated (arithmetic) mean and confidence limits for the lognormal model (see Section 3.2) and the non-parametric Kaplan-Meier method (see Section 3.5) are shown in the upper left area of Figure 1. The 95% confidence level is for a one-sided test as described in Section 3.2 (i.e. the interval is a 90% confidence interval for the mean). The estimated value of the 95% UCL for the 95th percentile $\hat{U}(0.95,0.95)$ based on the lognormal model (see Equation 12) is shown in the lower right of Figure 1 (see 95-95 Geometric TL). Also shown are the observed 95th percentile, the estimated 95th percentile from the lognormal model, and the value of R^2 . The estimates and confidence intervals can be obtained using

```
> lstats(ex1, 3000, 95, 95).
```

Prior to 1961 only selected workers ---see Watkins et al (1997), Watkins et al (2004) --- were monitored and for this example we assume that this individual worked in 1960 and was not monitored. An estimate of the unmonitored “missed dose” for each quarter in 1960 is needed. The doses are assumed to follow a lognormal distribution and the LOD is 0.3 mSv. That is, given the left censored sample from a lognormal distribution we want to estimate the prediction density for d , the unobserved quarterly doses in 1960. (see Section 3.4). The MLPD for $z = \log(d)$ is approximately normal with mean 3.013 and variance= $0.9917+(0.1706)^2$ --- see output from **mlnd2(ex1)**. The prediction density for the missed dose is lognormal with a geometric mean of 20.3 and geometric standard deviation of 2.775

4.2. EXAMPLE 2. BERYLLIUM EXPOSURE DATA

As part of a chronic disease prevention program, the Department of Energy (DOE) adopted a threshold limit value 8-hour time-weighted average (TWA) of 0.2 micrograms per cubic meter proposed by the American Conference of Government Industrial Hygienists (DOE 10 CFR Part 850). The development of the 8-hour occupational exposure limit for beryllium is discussed by Wambach and Tuggle (2000). 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 $\mu\text{g}/\text{m}^3$, i.e. this is an example of random (progressive) left censored data (available at the SAND web site in file Ex2.txt). The q-q plot in Figure 2 was obtained using the PLE as described in Section 3.5 using R function **plend(ex2)**. Both Figure 1 and

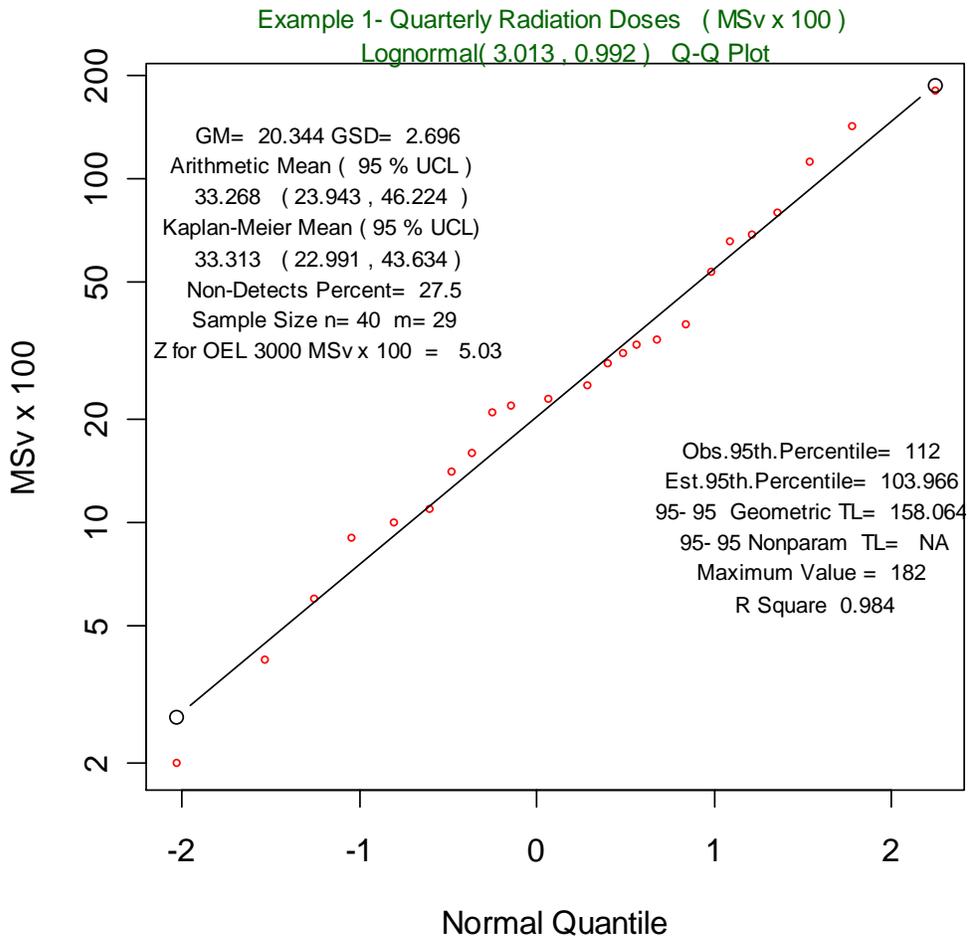


Figure 1. Lognormal Q-Q Plot for Example 1

Figure 2 can be obtained using R utility function `qqlogA()`. To obtain Figure 2, use the following at the R prompt:

```
>ex2<-read.table("Ex2.txt")
>qqlognA(ex2, "Example 2", OLE=0.2, unit = "mug/m^3").
```

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

```
> mlndln(ex2)
      mu          sigma      logE      sigma^2      -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 R function `mlndln()` is described in the Appendix and at the SAND website. To obtain the ML estimate of the 95-95 geometric upper tolerance limit (see Section 3.3 equation 12) calculate

$$\hat{y}_{.95} = \hat{\mu} + z_{.95} \hat{\sigma} = -2.652 \text{ and}$$

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

Then from equation 12 $\hat{U}(0.95,0.95) = \exp[-2.652 + 1.1659(0.0247)^{1/2}] = 0.091$.

4.3. EXAMPLE 3. LINEAR REGRESSION WITH NON-DETECTS

The data in Table 2 are the quarterly dose of record from 1956 to 1965 for a worker at the Oak Ridge Y-12 plant (see Example 1). The doses are assumed to follow the lognormal distribution with

$$E(y_i) = \mu_i = \alpha + \beta x_i$$

where $y_i = \log(d_i)$ and $x = \text{year} - 1961$, i.e. the intercept α represents the log dose in the first quarter of 1961 and β is the change per year in y . The dose data in Table 2 are in the first column of the file Ex3.txt at the SAND web site.

Table 2. Quarterly Film Badge Doses ⁺										
Year	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Q1	0	110	16	103	15	2	15	0	0	3
42	0	16	46	64	60	53	56	0	0	4
Q3	0	0	99	36	29	53	44	4	0	5
Q4	52	0	93	35	75	89	23	4	0	23
⁺ mSv*1008										

The 0s are changed to 30 (the LOD) and the censoring indicator is in column 2. Column 3 is the predictor variable $t61 = \text{year} - 1961$. ML estimates of α , β , and σ are obtained using the R driver function

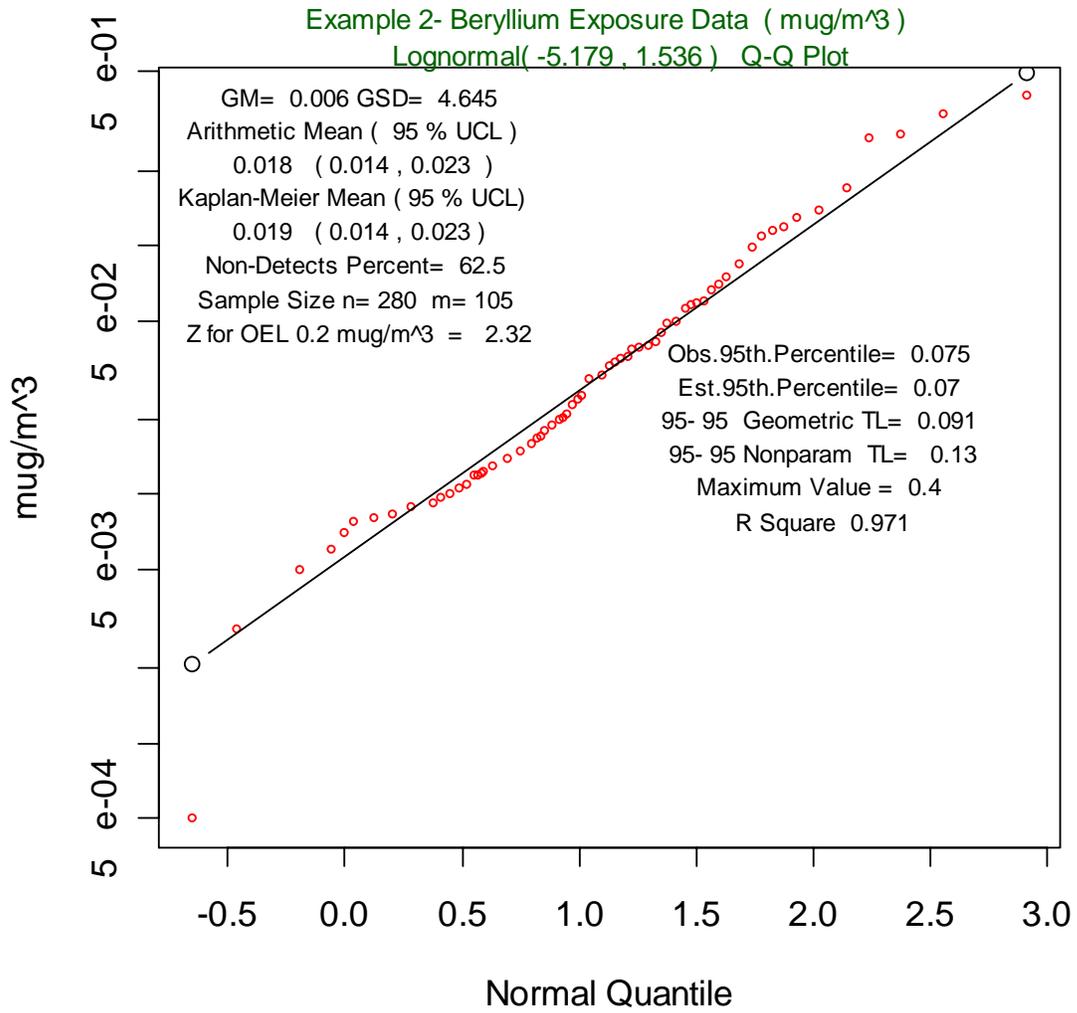


Figure 2. Lognormal Q-Q Plot for Example 2

`lnexh2()` shown in Exhibit 2 using the ML estimation method described in Section 3.4.

Exhibit 2 of Section 4.3

```
lnexh2 <- function(ww=ex3,lod=30)
{
# find ML estimates 1956-65 quarterly data
# data in ww Col 1 Col 2 Col 3
#      dose cen(0,1) t61
# y = log(dose) x = year - 1961
# E(y) = alpha + beta*x
#
# initial estimates using LS with zeros = lod/2
y0 <- log(ifelse( ww[,2]==0,lod/2,ww[,1] ))
go<- summary( lm( y0 ~ ww[,3] ) )
est <- c(go$coef[1,1], go$coef[2,1] , go$sigma )
names(est) <- c("alpha", "beta", "sigma")
# Use R function optim() with lognormal log-likelihood LNlr2()
# use Nelder-mead option to obtain ML estimates
opt <- optim(est,LNlr2, method = c("Nelder-Mead"),w=ww)
est <- opt$par
# use "L-BFGS-B" option to obtain estimate of var-covar matrix
opt <- optim(est,LNlr2,NULL, method = "L-BFGS-B",
  lower=c(-100.0,-100.0,0.1), upper=c(Inf,Inf,Inf),hessian=T,w=ww)
# use "L-BFGS-B" option again with pscale added
pscale <- c(sqrt(diag(solve(opt$hessian))))
opt <- optim(est,LNlr2,NULL, method = "L-BFGS-B",
  lower=c(-100.0,-100.0,0.1), upper=c(Inf,Inf,Inf),
  control=list(parscale=pscale), hessian=T,w=ww)
est <- opt$par
se <- sqrt(diag(solve(opt$hessian)))
vcvmle <- solve(opt$hessian) # ML covariance matrix
drc <- diag( 1/se ) ;
corm <- round( drc%*%vcvmle%*%drc,6) # correlation matrix
vcvmle <- round(vcvmle,9)
est <- c(opt$par,2*opt$value,opt$conver)
names(est) <- c("alpha", "beta", "sigma", "-2Log(L)", "Convrg")
vcv.cor <- cbind(vcvmle,corm)
se <- c(se,length(ww[,1]),NA)
mle <- rbind( est, se)
out <- list( mle ,vcvmle, corm )
names(out) <- list("MLE", "Variance-Covariance", "Correlation Matrices")
out
}

LNlr2 <- function(par=est,w)
{
# LNlr2 = - log liklihood for left censored sample lognormal
d<-w[,1]; cen<-w[,2] ; m<-par[1] + par[2]*w[,3] ;s<-par[3]
xx<-ifelse( cen==1,dlnorm(d,m,s,log=T) ,plnorm(d,m,s,log=T) )
-sum(xx)
}
```

Results obtained using R for the data in Table 2 are as follows:

```
> ex3 <- read.table("Ex3.txt",T)
> lnexh2(ex3,30)

$MLE
      α          β          σ      -2Log(L)  Convrge
est 3.0222162 -0.17484139 0.9906669 284.3878      0
se  0.1710291  0.06015249 0.1296071  40.0000     NA

$"Variance-Covariance"
      α          β          σ
α    0.029250951  0.000975915 -0.003819348
β    0.000975915  0.003618322  0.000136836
σ   -0.003819348  0.000136836  0.016797999
```

The ML estimates and standard errors for α , β , and σ are in the first two lines of output from `lnexh2(ex3)`, followed by the estimated variance-covariance matrix. This worker was not monitored prior to 1956 and the unmonitored dose in any quarter can be estimated using the prediction density (see equation 13). For example, for the first quarter in 1953 $x_f = 53 - 61 = -8$.

$$\begin{aligned} \hat{\mu}_f &= \hat{\alpha} + \hat{\beta} x_f = 3.022 - 0.1748(-8) = 4.421, & \text{and} \\ \text{var}(\hat{\mu}_f) &= \text{var}(\hat{\alpha}) + 2 x_f \text{cov}(\hat{\alpha}, \hat{\beta}) + x_f^2 \text{var}(\hat{\beta}) \\ &= 0.0293 + 2(-8)(0.00098) + (-8)^2(0.00362) = 0.245 \end{aligned}$$

The MLPD for $z_f = \log(d_f)$ is approximately normal with mean $\hat{\mu}_f$ and variance $\hat{\sigma}^2 + \text{var}(\hat{\mu}_f) = (.991)^2 + 0.245 = 1.227$. The MLPD for d_f (the first quarter of 1953) is approximately lognormal ($\hat{\mu}_f = 4.421$, $\hat{\sigma}_f = 1.227$) and the geometric mean is 83.2, the geometric standard deviation is 3.4, and the (arithmetic) mean is 176.6. The data from Table 2 are shown in Figure 3 along with the ML estimate (solid line) $E(y_t) = \hat{\mu}_t = \hat{\alpha} + \hat{\beta}(t-61)$. The x symbols (corresponding to non-detects) are obtained as the conditional expectation of y_t given that it is less than the log(LOD), i.e. $y_t^0 = \hat{\mu}_t - [n(z_t) / N(z_t)] \hat{\sigma}$, where $z_t = [\log(\text{LOD}) - \hat{\mu}_t] / \hat{\sigma}$, $n(z)$ is the standard normal density, and $N(z)$ is the normal CDF. This example illustrates how linear regression with non-detects can be used to estimate a workers dose during a quarter when the worker was not monitored. Watkins et al (2004) consider this problem in more detail and describe a better approach that uses data on a large group of workers with a similar employment and monitoring experience to describe the change in dose over time with a log-linear regression model. Groer and Ramachandran (2004) demonstrate the practical equivalence of ML estimation and Bayesian methods for these data.

1956-65 Quarterly Data for a Y-12 Worker

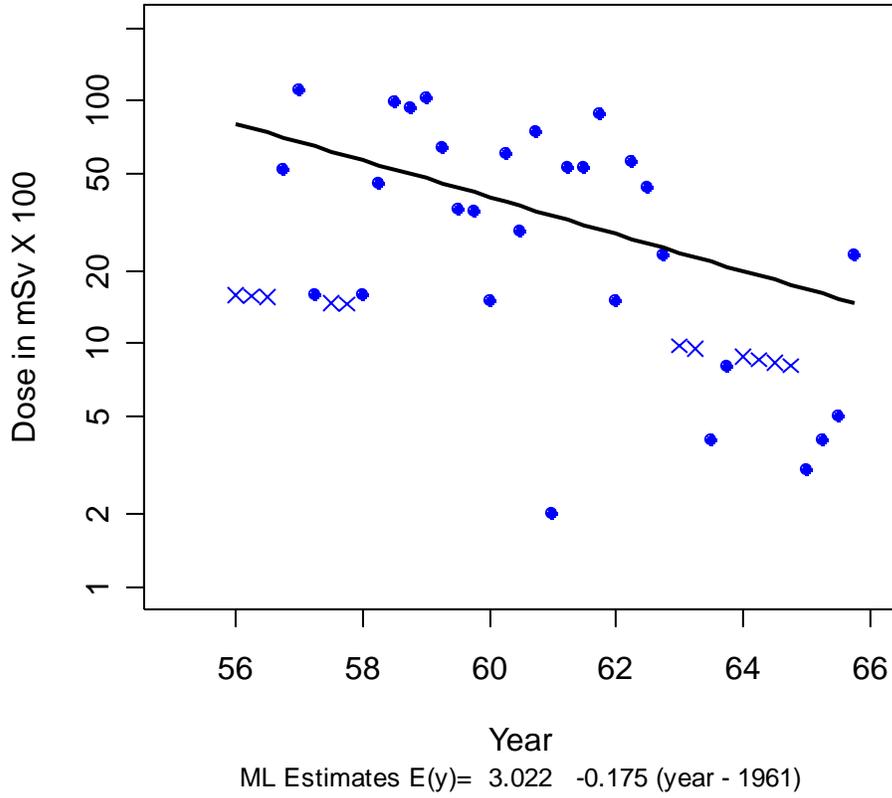


Figure 3. Data from Example 3

4.4. EXAMPLE 4. QUARTERLY GAMMA DOSES FOR A GROUP OF RADIATION WORKERS

Figure 4 shows the q-q plot and summary statistics for recorded doses for 844 workers at the Y-12 plant that were monitored for radiation exposure in the first quarter of 1959. Prior to 1961 only selected workers were monitored --- see Watkins et al (2004) for details.

Recorded doses are “true” doses (that have a between worker distribution) with “measurement error and recording policies” that result in recorded doses (many of which are non-detests). The recorded doses for a group of individuals during a quarter are described by a lognormal distribution. ML estimates of μ , σ , $\log(\mu_d)$, and σ^2 are obtained using R function `m1nd1n()` as follows:

```
> ex4<-read.table("Ex4.txt")
> m1nd1n(ex4)
```

	μ	σ	$\log E$	σ^2	$-2\text{Log}(L)$	Conver
mle	4.72692355	0.86932542	5.10483381	0.75573483	9.870420e+03	0
semle	0.03004886	0.02204226	0.03524518	0.03836954	-1.686256e-05	800

The 95-95 geometric UTL is 509.5, indicating compliance with the OEL limit, i.e. reject $H_0: D_p \geq 3000$ (see Section 2.2). These results can be used to estimate the dose for an unmonitored worker using the MLPD (Section 3.4). The MLPD $z = \log(d)$ will be normal with mean $\hat{\mu}$ and standard deviation $\hat{\sigma} = (\hat{\sigma}^2 + \text{var}(\hat{\mu}))^{1/2}$. This is equivalent to using equation 13 when there is no predictor variable (i.e. $p=1$) so that $\hat{\mu} = \hat{\alpha}$ and $\text{var}(\hat{\mu}) = \text{var}(\hat{\alpha})$. The MLPD for d is then approximately lognormal ($\hat{\mu} = 4.726$, $\hat{\sigma} = 0.8698$) with geometric mean 112.9, geometric standard deviation 2.4, and arithmetic mean 164.9.

5. DISCUSSION

The results in Sections 3 and 4 are based on large-sample 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 agreements 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.

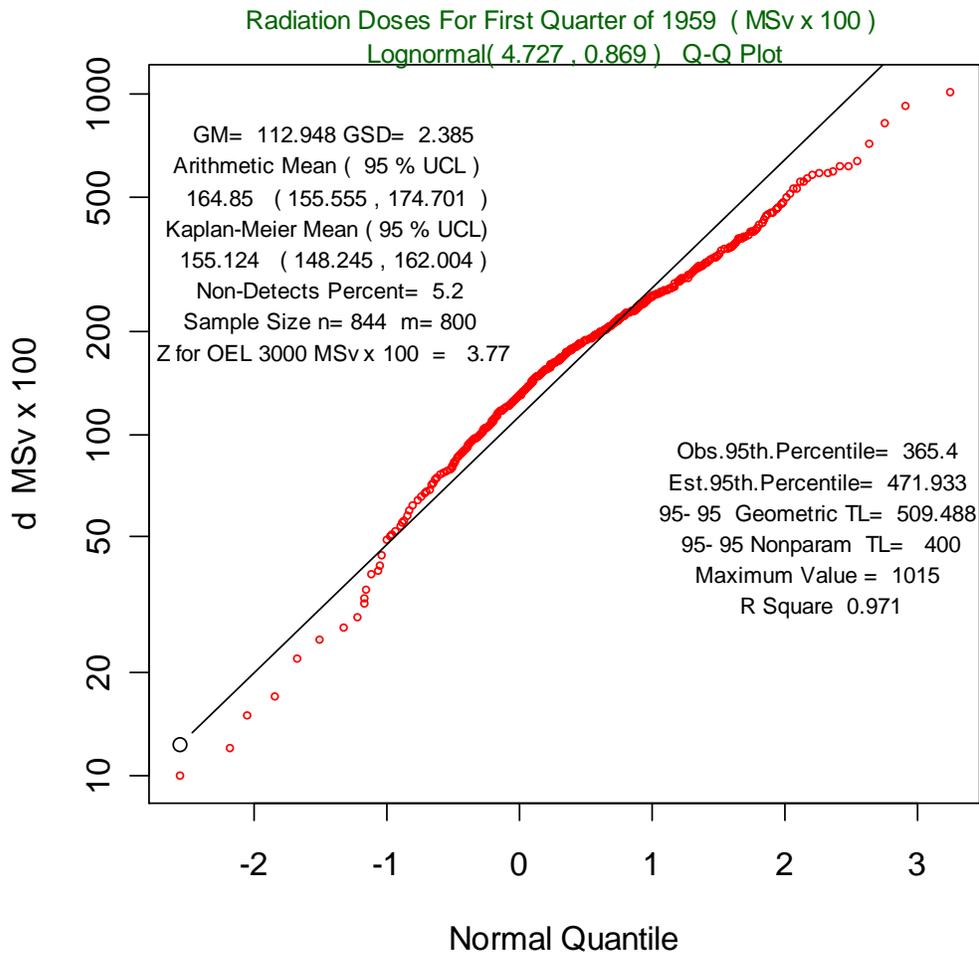


Figure 4. Lognormal Q-Q Plot for Example 4

In this report we have suggested that certain approximate methods for calculating confidence limits for functions of μ and σ may be useful as upper bounds for the large sample results. It also appears (based on limited simulation studies not reported here) that the percent non-detects will affect these limits. This issue will be considered in more detail in a subsequent report (Frome and Wambach, 2004).

Table 3 shows the results of applying method 1 (large sample ML) and method 2 (see Sections 3.2 and 3.3) to the data in examples 1, 2, and 4 to obtain upper confidence limits for μ_d and the 95th percentile (upper tolerance limit).

Table 3. 95 Percent Upper Confidence Limits						
μ			95 th Percentile			
Example	Method 1	Method 2	Method 1	Method 2	n	m
1	46.2	52.4	158.1	186.2	40	29
2	0.023	0.027	0.091	0.107	280	105
4	174.7	176.2	509.5	511.5	844	800

6. ACKNOWLEDGMENTS

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 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 Support activities to assist claimants and support the role of the Secretary of Health and Human Services under the Energy Employees Occupational Illness Compensation Program Act of 2000. The authors appreciate helpful comments and suggestions from the ORAU and NIOSH staff.

The authors thank the staff of ORAU's Center for Epidemiologic Research for support provided, in particular Jolene Jones for 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

- Armstrong, B. G., 1992. "Confidence Intervals for Arithmetic Means of Lognormally Distributed Exposures," *American Industrial Hygiene Association Journal*, 53(8), 481-485.
- Box, G.E.P. and G. C. Tiao, 1973. "Bayesian Inference in Statistical Analysis," Reading, MA: Addison-Wesley.
- Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, 1983. Graphical Methods for Data Analysis, Duxbury Press, Boston
- Cohen, A. C., 1991. "Truncated and Censored Samples," in *Theory and Applications*, Marcel Dekker, Inc., New York, Basel, Hong Kong.
- 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.
- Department of Energy, 10 CFR Part 850, Chronic Beryllium Disease Prevention Program, Federal Register, Vol. 64, No. 235, 68854-68914, December 1999.
- EEOICPA(2000), Energy Employees Occupational Illness Compensation Program Act (EEOICPA) of 2000, National Institute for Occupational Safety and Health (<http://www.cdc.gov/niosh/ocas/ocaseei.html>) Pub. L. 106-398, Title XXXVI, paragraph 3602
- Frome, E. L. and P. W. Wambach., 2004. "Analysis of Occupational Exposure Data with Non-Detectable Values," manuscript.
- Geisser, S., 1971. "The Inferential Use of Predictive Distributions," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart & Winston, 459-469.
- Groer, P.G. and R. Ramachandran, 2004. Bayesian Methods for Estimation of Missing Y-12 External Penetrating Doses with a Time Dependent Lognormal Model and for Imputation of an Individual's Missing Doses, ORAUT-TIB-0015 (in press)
- 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.
- Johnson, N. L. and B. L. Welch, 1940. "Application of the Non-Central t-Distribution," *Biometrika*, 31(3/4), 362-389.
- 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.

Lejeune, M. and G. D. Faulkenberry, 1982. "A Simple Predictive Density Function," *Journal of the American Statistical Association*, 77(379), 654-657.

Levy, M. S. and S. K. Perng, 1986. "An Optimal Prediction Function for the Normal Linear Model," *Journal of the American Statistical Association*, 81(393), 196-198.

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

Odeh, R. E. and D. B. Owen, 1980. Table for Normal Tolerance Limits, *Sampling Plans, and Screening*, Marcel Dekker, 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*, 599-601.

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, NY.

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.

Wambach, P. F. and R. M. Tuggle, 2000. "Development of an Eight-Hour Occupational Exposure Limit for Beryllium," *Applied Occupational and Environmental Hygiene*, 15(7), 581-587.

Watkins, J. P., D. L. Cragle, E. L. Frome, J. L. Reagan, C. M. West, D. Crawford-Brown and W. G. Tankersley, 1997. "Collection, Validation, and Treatment of Data for a Mortality Study of Nuclear Industry Workers," *Applied Occupational and Environmental Hygiene*, 12(3), 195-205.

Watkins, J. P., G. D. Kerr, E. L. Frome, W. G. Tankersley and C. M. West, 2004. *Historical Evaluation of the Film Badge Program at the Y-12 Facility in Oak Ridge, Tennessee: Part I – Gamma Radiation*, ORAU Technical Report #2004-0888, Oak Ridge Associated Universities, Oak Ridge, TN.

APPENDIX

R (2004) is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code and binary 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/> (see e.g. An Introduction to R under the “Manuals” link.) Additional manuals, tutorials, etc. are provided by users of R under the “Contributed” Link. Additional references are provide under the “Publications” link and the book by Venables and Ripley(2002) is highly recommended.

All of the R functions discussed in this report and the data used in the examples in Section 4 are available at the Statistical Analysis of Non-Detects (SAND) website at URL <http://www.csm.ornl.gov/~frome/sand>

Most of the serious computing is done by R base functions **optim()** and **uniroot()**. The R functions described in this report are provided to assist the reader that may not have experience with R. They are not “formal” R functions, i.e. there is no error checking or online “help” files. Documentation for each function is provided in this report and as comments in each function. All of the files at the SAND 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 main.R (see Exhibit 3). The additional functions reflect the authors’ interest and require revisions for other applications. They are also provided in the file util.R at the SAND website.

Exhibit 3 in the Appendix

```
# Listing of R functions:
#
# mlndln() calculates ML estimates for left censored sample
# extol() exact tolerance limit for Lognormal model
# nptl() calculate index for Nonparametric tolerance limit
#
##### mlndln #####
mlndln <- function(dd = ex1 )
{
# ML estimates for lognormal sample with non-detects
# see ORNL/TM-2004/146 Section 3
# USAGE: mlndln( dd )
# ARGUMENT: matrix dd with d[i] in column 1 and cen[i] in col 2
# d[i] is positive lognormal data cen[i]=0 for non-detect ; 1 for detect
# y= log(d) is normal with mean mu and standard deviation sigma
# E(dose) = exp( mu + 0.5*sig2)= exp(logE) where sig2 = sigma^2
# m is number of detects and Conver is convergence check
# VALUE: mlndln returns 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 function 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 )$par
cont <- list(parscale=abs(est))
opt1 <- optim(est,ndln ,NULL, method = "L-BFGS-B",lower=c(-Inf,0.0),
upper=c(Inf,Inf),cont, hessian=T,xd=dd )
conv1 <- opt1$conver # 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
# est logE(dose) and sig2 (sigma^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$conver+conv1 )
semle <- c(semle,sqrt(diag(solve(opt2$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 liklihood for lognormal sample
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)
{
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)
}

##### extol #####
extol <- function(n=50,p=0.95,gam=0.95)
{
# For random sample size n from normal distribution
# ybar is sample mean and SD is standard deviation
# calculate with confidence level gam that at least
# 100p percent of population lies below the
# tolerance limit = ybar + k*SD
# USAGE: extol(n,p,gam)
# ARGUMENTS: n: sample size p: defined above
# gam: 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
# REFERENCES:
# Johnson, N. L. and Welch, B. L. (1940), Applications
# of the Non-Central T distribution, Biometrika, 362-389
# see Table 1 in
# Odeh, R.E. and Owen, D.B.(1980) Table for Normal Tolerance Limits,
# Sampling Plans, and Screening, Marcel Dekker, New York
# NOTE: second argument to uniroot may not be optimal

tx <- function(x,nn=n,th=p,ga=gam)
{pt(x,nn-1,(-sqrt(nn)*qnorm(th))) + ga- 1}

uout <- uniroot(tx,sqrt(n)*c( -(1/(1- max(p,gam) )),50) )
u.tmp <- uout
k<- -uout$root/sqrt(n)
k
}

##### nptl #####
nptl <- function(n=100,p=0.95,gamma=0.95)
{
# function nptl(n,p,gam) given n p and gamma 8Oct2002
# 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
# USAGE: nptl(n,p,gam)
# ARGUMENTS: n: sample size p: defined above
# gam: confidence level for one-sided interval
# VALUE: m
# DETAILS: Requires base R function qbeta(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
}

```

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. D. Cragle (CragleD@ornl.gov)
9. P. Groer (groer@utk.edu)
10. J. Kenoyer (jkenoyer@ornl.gov)
11. G. Kerr (gdkerr@bellsouth.net)
12. R. Toohey (TooheyR@ornl.gov)
13. T. Taulbee (tgt4@cdc.gov)
14. W. Tankersley (TankersB@ornl.gov)
15. P. Wambach (Paul.Wambach@eh.doe.gov)
16. J. Watkins (WatkinsJ@ornl.gov)