United States
Environmental Protection
Agency

Office of
Research and
Development

Office of Solid
Waste and
Emergency
Response

EPA/600/S-97/006
December 1997

Technology Support Center Issue

The Lognormal Distribution in Environmental Applications

Ashok K. Singh[1], Anita Singh[2], and Max Engelhardt[3]

The Technology Support Projects, Technology Support Center (TSC) for Monitoring and Site Characterization was established in 1987 as a result of an agreement between the Office of Research and Development (ORD), the Office of Solid Waste and Emergency Response (OSWER) and all ten Regional Offices. The objectives of the Technology Support Project and the TSC were to make available and provide ORD's state-of-the-science contaminant characterization technologies and expertise to Regional staff, facilitate the evaluation and application of site characterization technologies at Superfund and RCRA sites, and to improve communications between Regions and ORD Laboratories. The TSC identified a need to provide federal, state, and private environmental scientists working on hazardous waste sites with a technical issue paper that identifies data assessment applications that can be implemented to better define and identify the distribution of hazardous waste site contaminants. The examples given in this Issue paper and the recommendations provided were the result of numerous data assessment approaches performed by the TSC at hazardous waste sites. Mr. John Nocerino provided guidance and suggestions that greatly enhanced the quality of this Issue Paper.

This paper was prepared by A. K. Singh, A. Singh, and M. Engelhardt. Support for this project was provided by the EPA National Exposure Research Laboratory's Environmental Sciences Division with the assistance of the Superfund Technical Support Projects Technology Support Center for Monitoring and Site Characterization, OSWER’s Technology Innovation Office, the U.S. DOE Idaho National Engineering and Environmental Laboratory, and the Associated Western Universities Faculty Fellowship Program. For further information, contact Ken Brown, Technology Support Center Director, at (702) 798-2270, A. K. Singh at (702) 895-0364, A. Singh at (702) 897-3234, or M. Engelhardt at (208) 526-2100.

Purpose and Scope

The purpose of this issue paper is to provide guidance to environmental scientists regarding the interpretation and statistical assessment of data collected from sites contaminated with inorganic and organic contaminants. Contaminant concentration data from sites quite often appear to follow a skewed probability distribution. The lognormal distribution is frequently used to model positively skewed contaminant concentration distributions. The H-statistic

[1] Department of Mathematics, University of Nevada, Las Vegas, NV 89154
[2] Lockheed Martin Environmental Systems & Technologies, 980 Kelly Johnson Dr., Las Vegas, NV 89119
[3] Lockheed Martin Idaho Technologies, P.O. Box 1625, Idaho Falls, ID 83415-3730

COMMUNICATION • TRAINING • TECHNOLOGY SUPPORT •
Technology
Support
Project

Technology Support Center for
Monitoring and Site Characterization
National Exposure Research Laboratory
Environmental Sciences Division
Las Vegas, NV 89193-3478

Technology Innovation Office
Office of Solid Waste and Emergency Response,
U.S. EPA, Washington, D.C.

Walter W. Kovalick, Jr., Ph.D., Director

Printed on Recycled Paper

based Upper Confidence Limit (UCL) for the mean of a lognormal population is recommended by U.S. EPA guidance documents (see, for example, EPA (1992)) and is widely used to make remediation decisions at Superfund sites. However, recent work in environmental statistics has cast some doubts on the performance of the formula based on the H-statistic for computing an upper confidence limit of the mean of a lognormal population. This issue paper is mainly concerned with the problem of computing an upper confidence limit when the contaminant concentration distribution appears to be highly skewed.

Several approaches to computing upper confidence limits for the mean of a lognormal population are considered. The approaches discussed include those based on the H-statistic, the jackknife method, the bootstrap method, and a method based on the Chebychev inequality. Simulated examples show that for values of the coefficient of variation larger than 1, the upper confidence limits for the mean contaminant concentration based on the H-statistic are much higher than the upper confidence limits obtained by the other estimation methods. This may result in an unnecessary cleanup. In other words, the use of the jackknife method, the bootstrap method, or the Chebychev inequality method provides better input to the risk assessors and may result in a significant reduction in remediation costs. This is especially true when the number of samples is thirty or less. When the value of the coefficient of variation exceeds 1, upper confidence limits based on any of the other estimation procedures appear to be more stable and reliable than those based on the H-statistic. Values of the coefficient of variation computed from observed contaminant concentrations are typically used by environmental scientists to assess the normality of the population distribution. In this issue paper, the issue of using the coefficient of variation in environmental data analysis is addressed and the problem of estimating the coefficient of variation, when sampling from lognormal populations, is also discussed.

This issue paper is divided into the following major sections: (1) Introduction, (2) the Lognormal Distribution, (3) Methods of Computing a UCL of the Mean, (4) Examples, and (5) Summary and Recommendations.

1. Introduction

Most of the procedures available in the literature of environmental statistics for computing UCL of the mean of a population assume that contaminant concentration data is approximately normally distributed. However, the distributions of contaminant concentration data from Superfund sites typically are positively skewed and are usually modeled by the lognormal distribution. This apparent skewness, however, may be due to biased sampling, multiple populations, or outliers, and not necessarily due to lognormally distributed data.

Biased sampling is often used in sampling for site characterization (Power, 1992). Another common situation often present with environmental data is a mixed distribution of several subpopulations (see Figure 1). Also, the presence of one or more outliers, spurious observations, or anomalies can result in a data set which appears to come from a highly skewed distribution. When dealing with a skewed distribution, statisticians sometimes recommend

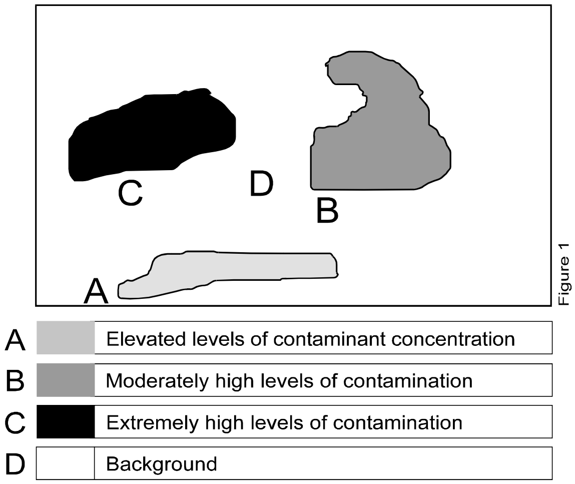
C
D
B
A
Figure 1
A
Elevated levels of contaminant concentration
B
Moderately high levels of contamination
C
Extremely high levels of contamination
D
Background

Figure 1 A site with several sources of contamination.

using the population median (instead of the population mean) as a measure of central tendency. However, remediation decisions at a polluted site typically are made on the basis of the population mean, and therefore UCL of the mean of the concentration distribution is needed. For positively skewed distributions, the median is smaller than the mean: therefore a UCL for the median provides an inappropriate basis for a decision about the mean.

U.S. EPA guidance documents recommend the use of H-statistics to compute the UCL of the mean of a lognormal distribution (EPA, 1992). A detailed discussion of H-statistics is given in Gilbert (1987). For data sets with nondetects, estimation methods developed for censored data from a lognormal distribution are discussed by Lecher (1991). The use of the lognormal distribution has been controversial because it can lead to incorrect decisions. For example, recent work of Gilbert (1993) indicates that statistical tests of hypotheses based on H-statistics can yield unusually high false positives, which would result in an unnecessary cleanup. The situation may be reversed when dealing with estimation of the mean background level. If the H-statistic based method is used to compute a UCL of the mean for the observed background concentrations, then the mean of the background level may be overestimated, which may result in not remediating a contaminated area of the site. Stewart (1994) also showed that the incorrect usage of a lognormal distribution may lead to erroneous results.

Most of the "classical" statistical methods based on the normal distribution were developed between 1920 and 1950 and have been well investigated in the statistical literature. On the other hand, lognormal-based methods have not received the same level of scrutiny. Furthermore, the classical methods became popular due to their computational convenience. The 1980s have produced a new breed of statistical methods based on the power and availability of computers (see, for example, Efron and Gong, 1983). Both the jackknife and bootstrap methods require a great deal of computer power, and, therefore, have not been widely adopted by environmental statisticians. However, with the recent advances in computer equipment and software, computationally intensive statistical procedures have become more practical and accessible.

The authors of this article have critically reviewed several estimation procedures which can be used to compute UCL values via monte carlo simulation. These include the simple arithmetic mean, the Minimum Variance Unbiased Estimate (MVUE), and nonparametric procedures such as the jackknife and the bootstrap procedures. Computer simulation experiments (not included in this paper) have been performed for various values of the population standard deviation, or equivalently the Coefficient of Variation (CV), and sample sizes ranging from 10 to 101. It has been demonstrated that for samples of size 30 or less, the H-statistic based UCL results in unacceptably high estimates of the threshold levels such as the background level contamination. This is especially true for data sets from populations with CV values exceeding 1. For samples of larger sizes, the use of H-statistics can be replaced by UCLs based on nonparametric methods such as the jackknife or the bootstrap. Other well-known results such as the central limit (CLT) and Chebychev theorems may also be used to obtain UCLs. To illustrate problems associated with methods based on lognormal theory, results for some simulated examples and some from Superfund work done by the authors have been included in this paper.

2. The Lognormal Distribution

The authors briefly describe the lognormal distribution. By definition, contaminant concentration is lognormally distributed if the log-transformed concentrations are normally distributed. This can be mathematically described as follows:

If $Y = \ln(X)$ is normally distributed with mean, μ, and variance, σ^2, then X is said to be lognormally distributed with parameters μ and σ^2. It should be noted that μ and σ^2 are not the mean and variance of the lognormal random variable, X, but they are the mean and variance of the log-transformed random variable, Y. However, it is common practice to use the same parameters to specify either, and it is convenient to refer to the normal distribution with the abbreviated notation $Y \sim \mathrm{N}(\mu, \sigma^2)$ and the log-normal distribution with the abbreviation $X \sim \mathrm{LN}(\mu, \sigma^2)$. Figure 2, which shows plots of a normal and a lognormal density function with $\mu = 0$ and $\sigma^2 = 0.5$, illustrates the difference between normal and lognormal distributions.

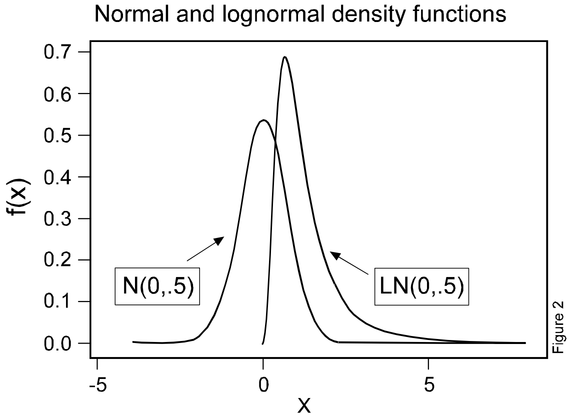

Figure 2 Graphs of normal N(μ = 0, σ^2 = 0.5) and lognormal LN(μ = 0, σ^2 = 0.5) density functions.

Figure 3, which shows plots of several lognormal distributions, each with μ = 0, illustrates how varying the parameter σ^2 can change the amount of skewness.

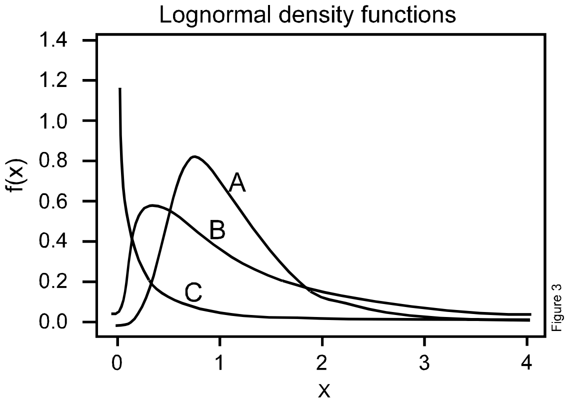

Figure 3 Graphs of A: LN(μ = 0, σ^2 = 0.25), B: LN(μ = 0, σ^2 = 1.0) and C: LN(μ = 0, σ^2 = 25.0) density functions.

The parameters of interest of a lognormal distribution, LN(μ, σ^2), are given as follows:

$$\text{Mean} = \mu_1 = \exp(\mu + 0.5\sigma^2) \tag{1}$$

$$\text{Median} = \exp(\mu) \tag{2}$$

$$\text{Variance} = \sigma_1^2 = [\exp(2\mu + \sigma^2)]\,[\exp(\sigma^2) - 1] \tag{3}$$

$$\text{Coefficient of Variation} = CV = \sigma_1/\mu_I = \sqrt{\exp(\sigma^2) - 1} \tag{4}$$

$$\text{Skewness} = (CV)^3 + 3(CV). \tag{5}$$

Throughout this paper, irrespective of the underlying distribution, μ_I, and σ_I^2 represent the mean and variance of the random variable X (in original units), whereas μ and σ^2 are the mean and variance of its logarithm given by Y=ln(X). The pth quantile (or 100pth percentile), x_p, of the distribution of a random variable, X, is defined by the probability statement $P(X \le x_p) = p$. If z_p is the pth quantile of the distribution of the standard normal random variable, Z, with $P(Z \le z_p) = p$, then the pth quantile of a lognormal distribution is given by $x_p = \exp(\mu + z_p\sigma)$. For example, on the average, 95% of the observations from a lognormal LN(μ, σ^2) distribution would lie below $\exp(\mu + 1.65\sigma)$. Because the 0.5th quantile of the standard normal distribution is $z_{0.5} = 0$, the 0.5th quantile (or median) of a lognormal distribution is $\exp(\mu)$, which is obviously smaller than the mean, μ_1, which is given by equation (1). In this paper, several procedures to estimate the UCL of the mean have been considered. Ordinarily, one would expect the spread of an estimate of the mean to be smaller than the spread of the population itself (see Figure 4). Thus, intuitively, the 95% UCL of the mean should be smaller than the 95th percentile of the corresponding lognormal distribution. In many instances with lognormal-based methods, this statement is violated even on lognormal data, especially for smaller sample sizes. The quantiles discussed above are used later to shed some light on the behavior of the UCL of the mean which are based on the H-statistic.

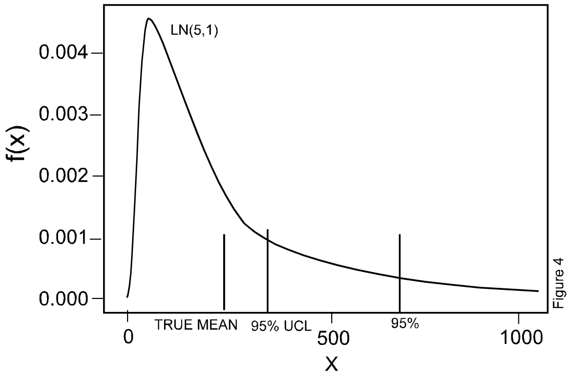

Figure 4 Graphs showing the relative positions of the TRUE MEAN, the 95% UCL, and the 95th percentile.

One of the inherent assumptions required to compute the UCL of the mean is that the data set under consideration comes from a single statistical population (e.g., background only). Violation ofthis assumption can lead to invalid applications of a statistical technique. The consequences of this assumption being violated are discussed as follows. A data set can be put into a statistical procedure (e.g., the Shapiro-Wilks test of normality) or a computer program whether or not the required assumptions are met. It is the user's responsibility to ensure the underlying assumptions required to conduct the statistical procedure are met. The decisions and conclusions derived from incorrectly used statistics can be expensive. For example, incorrect use of a statistic may lead to wrong conclusions such as: 1) remediation of a clean part of the site, or 2) no remediation at a contaminated part of the site. The first wrong conclusion will result in an unnecessary cleanup whereas the second incorrect conclusion may cause a threat to human health and the environment. It is likely that the availability of new and improved statistical software has also increased the misuse of statistical techniques. This is illustrated in the following discussion of the application to some simulated and real data sets. It should be reiterated that it is the analyst's (user's) responsibility to verify that none of the required assumptions are violated before using a statistical test and deriving inferences based upon the resulting analysis. In many cases, this may warrant expert advice from a statistician.

Often, the central portion of a data set will behave as if it came from a normal distribution. However, in practice, a normally distributed data set with a few extreme (high) observations can be incorrectly modeled by the lognormal distribution with the lognormal assumption hiding the outliers. Also, the mixture of two or more normally distributed data sets with significantly different mean concentrations such as one coming from the clean background part and the other taken from a contaminated part of the site can also be modeled (incorrectly) by a lognormal distribution. The following example illustrates this point.

Example 2.1. Simulated data set from two populations

A simulated data set of size fifteen (15) has been obtained from a mixture of two normal populations. Ten observations (representing background) were generated from a normal distribution with mean, 100, and standard deviation, 50, and five observations (representing contamination) were generated from a normal distribution with mean, 1000, and standard deviation, 100. The mean of this mixture distribution is 400. The generated data are as follows: 180.5071, 2.3345, 48.6651, 187.0732, 120.2125, 87.9587, 136.7528, 24.4667, 82.2324, 128.3839, 850.9105, 1041.7277, 901.9182, 1027.1841, and 1229.9384.

Discussion of Example 2.1

The data set in Example 2.1 failed the normality test based on several goodness-of-fit tests such as the Shapiro-Wilks, W-test (W=0.7572), and Kolmogorov-Smirnov (K-S = 0.35) tests (see Figures 5 and 6). However, when these tests were carried out on the log-transformed data, the test statistics are insignificant at the $\alpha = 0.05$ level of significance with W=0.8957, and K-S = 0.168, suggesting that a lognormal distribution (see Figures 7 and 8) provides a reasonable fit to the data. Based upon this test, one might incorrectly conclude that the observed concentrations come from a single background lognormal population. This incorrect conclusion is made quite frequently. This data set is used later to illustrate how modeling the mixture data set by a lognormal distribution will result in incorrect estimates of mean contamination levels at various parts of the site.

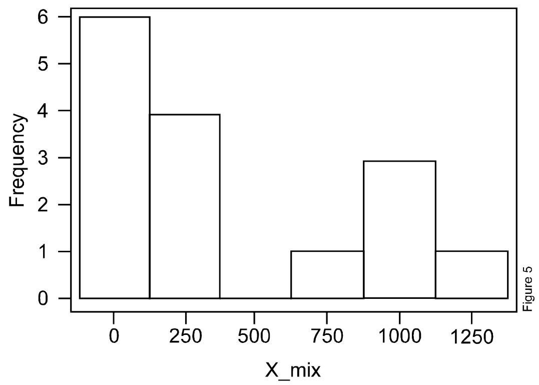

Figure 5 Histogram of the 15 observations from the mixture population of Example 2.1.

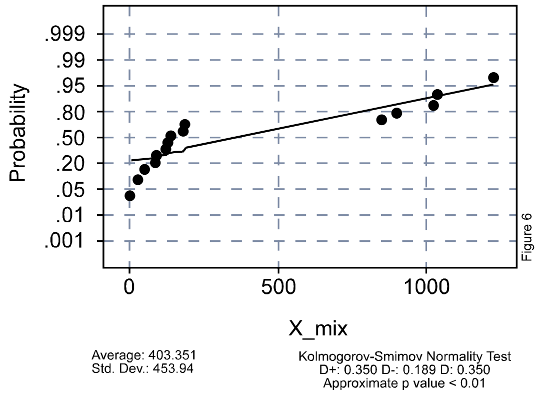

Figure 6 K-S test of normality for the data of Example 2.1.

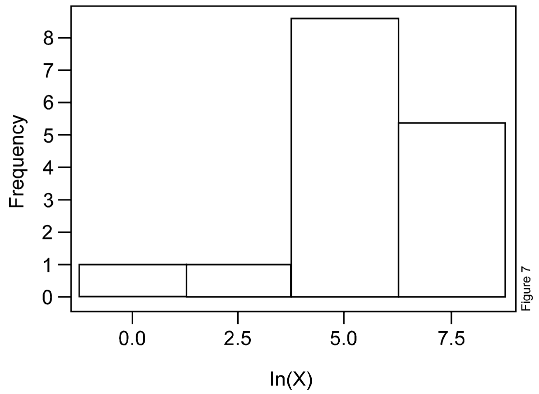

Figure 7 Histogram of the log-transformed 15 observations from the mixture population of Example 2.1.

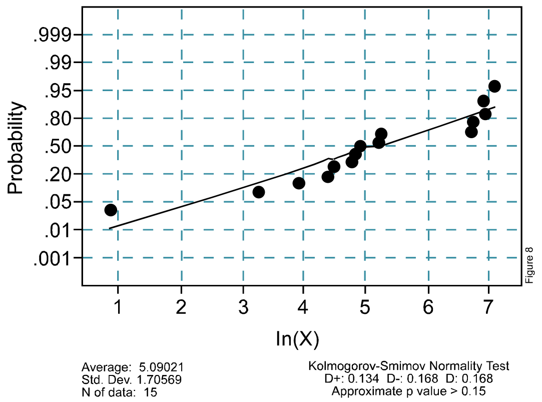

Figure 8 K-S test of lognormality for the data of Example 2.1.

3. Methods of Computing a UCL of the Mean

The main objective of this study is to assess the performances of the various methods of estimating the UCL for the mean, μ_1, of positively skewed populations. The assumption of a lognormal distribution to model such populations has become quite popular among environmental scientists (Ott, 1990). As noted in Section 2, for positively skewed data sets, there are potential problems in using standard methods based on the lognormal theory. Therefore, we will compare the lognormal-based methods often used with cleanup standards with some other available methods. The alternate methods considered here have the advantage that they do not require assumptions about the specific form of the population distribution. In other words, they do not assume normality or lognormality of the data set under consideration. In Section 4, the UCL of the mean has been computed for several examples using the following methods:

- The H-statistic
- The Jackknife procedure
- The Bootstrap procedure
- The Central Limit Theorem
- The Chebychev Theorem

A brief description of the computation of the various estimates and the associated confidence limits obtained using the above-mentioned procedures follows:

Parametric Lognormal Procedures

Let $x_1, x_2, \ldots, x_n$ be a random sample from a lognormal distribution with mean, μ_1, and variance, σ_1^2, and denote by μ and σ the population mean and population standard deviation (sd), and $\bar{y}$, and s_y the sample mean and sample sd, respectively, of the log-transformed data $y_i = \ln(x_i)$; $i = 1, 2, \ldots, n$. . Specifically,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6}$$

and

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{7}$$

In a more general setting, consider a population with an unknown parameter, θ. The minimum variance unbiased estimate (MVUE) of θ is the one that is not only an unbiased estimate of θ (i.e., the expected value of the estimate is equal to the true value of the parameter), but it also has a smaller variance than any other unbiased estimate of θ. When the parameter of interest is the mean, μ_1, of a lognormally distributed population, Bradu and Mundlak (1970) derive its MVUE, which is given by

$$\hat{\mu}_1 = \exp(\bar{y})\, g_n(s_y^2/2) \tag{8}$$

where $g_n(u)$ is a function whose form is rather complicated, but an infinite series solution is given by Aitchison and Brown (1976). Tabulations of this function are provided by Gilbert (1987, Table A9). Note that Gilbert uses ψ_n in place of g_n. This function is also used in computing the MVUE of the variance, σ_1^2, of a lognormal population, as given by Finney (1941),

$$\hat{\sigma}_1^2 = \exp(2\bar{y})[g_n(2s_y^2) - g_n((n-2)s_y^2/(n-1))]. \tag{9}$$

Bradu and Mundlak (1970) give the MVUE of the variance of the estimate $\hat{\mu}_1$,

$$\hat{\sigma}^2(\hat{\mu}_1) = \exp(2\bar{y})[[g_n(s_y^2/2)]^2 - g_n((n-2)s_y^2/(n-1))]. \tag{10}$$

Another estimate which is also sometimes used is known as the Maximum Likelihood Estimate (MLE). When the data set is a random sample from a lognormal distribution, the MLE of the parameter, μ, is simply the sample mean of the log-transformed data, $\hat{\mu} = \bar{y}$, and the MLE of σ^2 is a multiple of the sample variance of the log-transformed data, namely, $\hat{\sigma}^2 = [(n-1)/n]s_y^2$. The MLE of any function of the parameters μ and σ^2 is obtained by simply substituting these MLEs in place of the parameters. For example, the MLE of the mean of a lognormal population is $\exp(\hat{\mu} + 0.5\hat{\sigma}^2)$, and the MLE of the 95th percentile is $\exp(\hat{\mu} + 1.65\hat{\sigma})$. One disadvantage of the MLEs for the lognormal mean and percentiles is that they are biased estimates. Another slight modification uses s_y in place of the MLE, $\hat{\sigma}$. Although the result is not identical to the MLE, there is only a small difference numerically, and for convenience the use of the term MLE will also include this modified version.

Finally, the one-sided $(1-\alpha)100\%$ UCL for the mean, μ_1, of the lognormal distribution derived by Land (1971, 1975) is given as follows:

$$UCL = \exp[\bar{y} + 0.5s_y^2 + s_y H_{1-\alpha}/\sqrt{n-1}]. \tag{11}$$

Tables of H-statistic values can be found in Land (1975) and also in Gilbert (1987, Table A10).

Use of the UCL for a population mean based on the H-statistic is widely recommended in environmental guidance documents. Theoretically, the UCL based on the H-statistic has optimal properties when the population is truly lognormal. However, in practice the results can be quite disappointing and misleading if the data set includes outliers, or is a mixture of data from two or more distributions. Monte carlo investigations performed by the authors confirm that, for small sample sizes, the use of the H-statistic approach can result in unacceptably high values of UCL when the CV is larger than 1.0. Consequently, other methods for computing a UCL of the mean, μ_1, of a distribution of unspecified form will be considered and the results compared with UCLs obtained by the H-statistic approach.

The methods considered in this paper can be viewed as variations of a basic approach to constructing confidence intervals known as the pivotal quantity method. In general, a pivotal quantity is a function of both the parameter θ and an estimate $\hat{\theta}$ such that probability distribution of the pivotal quantity does not depend on θ. Perhaps the best-known example of a pivotal quantity is the well-known t statistic,

$$t = \frac{\bar{x} - \mu_1}{s_x/\sqrt{n}} \tag{12}$$

where $\bar{x}$ and s_x are, respectively, the sample mean and sample standard deviation. If the data is a random sample from a normal population with mean, μ_1, and standard deviation, σ_1, then the distribution of this pivotal quantity is the familiar Student's t distribution with $n-1$ degrees of freedom. Because the Student's t distribution does not depend on either unknown parameter, quantiles are available. Denote by $t_{\alpha,n-1}$ the upper αth quantile of the Student's t distribution with $n-1$ degrees of freedom. Based on equation (12), it is possible to derive a $(1-2\alpha)100\%$ confidence interval of the form

$$\left(\bar{x} - t_{\alpha,n-1}s_x/\sqrt{n},\ \bar{x} + t_{\alpha,n-1}s_x/\sqrt{n}\right). \tag{13}$$

The confidence interval is given in the familiar form of a two-sided confidence interval for the mean. If the lower limit of this interval is disregarded, the upper limit provides a $(1-\alpha)100\%$ UCL for the mean, μ_1.

For a population which is normally distributed, equation (13) provides the best way of constructing confidence intervals for the population mean. However, as noted previously, the distribution of contaminant concentration data is typically positively skewed and frequently involves outliers. It is well known that the sample mean and sample standard deviation get severely distorted in the presence of outliers, (Singh and Nocerino 1995), and consequently any function, such as the Student's t, given by equation (12) above of these statistics also gets severely influenced by the presence of outliers. Robust methods for estimating the population mean and sd are available in the software package, SCOUT, as identified in Singh and Nocerino (1995). In practice, statistical procedures based on the pivotal quantity equation (12) are usually thought to be "robust" relative to violation of the normality assumption. As noted by Staudte and Sheather (1990), tests based on the Student's t are nonrobust in the presence of outliers. Consequently, other procedures which do not rely on a specific parametric assumption for the population distribution are also considered in the following discussion.

The approach of constructing confidence intervals from pivotal quantities (or approximate pivotal quantities) permits a unified treatment of these alternate procedures. In particular, each procedure involves an approximate pivotal quantity with the difference between the unknown population mean, μ_1, and a point estimate of the mean in the numerator, and an estimate of the standard error of the point estimate in the denominator. Thus, each procedure involves two parts: 1) finding some reasonably robust estimate of the mean, (Singh and Nocerino 1995), and 2) providing a convenient way to obtain quantiles of the pivotal quantity. A general discussion of the pivotal quantity approach to constructing confidence intervals is given by Bain and Engelhardt (1992).

As noted above, in order to apply the pivotal quantity method, it is necessary to have quantiles of the distribution of the pivotal quantity. For example, in order to compute equation (13), it is necessary to have quantiles of the Student's t distribution. These quantiles can be found in tables or computed with the appropriate software. However, for nonnormal populations the required quantiles are not, in general, readily available. In some cases, even though the exact distribution of a pivotal quantity is not known, an approximate distribution can be used. Thus, except for the H-statistic approach, which is exact if the population is truly lognormal, all of the other methods discussed below give only approximate UCL values for the population mean. The true confidence level of UCLs will vary from one method to the next, and without some additional study, it will not be clear whether the comparisons are fair. In other words, it is possible to have a smaller UCL at the expense of a true confidence level which is below the nominal level, and below the true confidence level of another competing method.

In environmental applications, the objectives typically are: 1) the identification of hot spots, which are typically represented by the high extreme concentrations, or 2) the separation of clean part(s) of a site from the dirty contaminated part(s) of the site. However, from the examples discussed in the following, it can be seen that the practical use of the lognormal distribution in those environmental applications is questionable as a lognormal distribution often accommodates extreme outlying observations and mixture populations as part of one lognormal distribution.

Jackknife and Bootstrap Procedures

General methods for deriving estimates, such as the method of maximum likelihood, often result in estimates which are biased. Bootstrap and jackknife procedures as discussed by Efron (1982) and Miller (1974) are nonparametric statistical techniques which can be used to reduce the bias of point estimates and construct approximate confidence intervals for parameters such as the population mean. These two procedures require no assumptions regarding the statistical distribution (e.g., normal or lognormal) for the underlying population, and can be applied to a variety of situations no matter how complicated. However, it should be pointed out that a use of a parametric statistical method (depending upon distributional assumptions), when appropriate, is more efficient than its nonparametric counterpart. In practice, parametric assumptions are often difficult to justify, especially in environmental applications. In these cases, nonparametric methods are valuable tools for obtaining reliable estimates of the parameters of interest. Although bootstrap and jackknife procedures are conceptually simple, they are based on resampling techniques requiring considerable computing power and time.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with an unknown parameter θ (***e.g.***, $\boldsymbol{\theta = \mu_1}$), and let $\hat{\theta}$ be an estimate of θ which is a function of all n observations. For example, the parameter θ could be the mean, and a reasonable choice for the estimate $\hat{\theta}$ might be the sample mean, $\overline{X}$. Another choice is the MVUE of a lognormal mean. Of course, if the population is not lognormal then this estimate may not perform well: but, because it is frequently used with skewed data sets, it is of interest to see how it performs relative to the other methods.

Jackknife Estimation

In the jackknife approach, n estimates of θ are computed by deleting one observation at a time. Specifically, for each index, i, denote by $\hat{\theta}_{(i)}$ the estimate of θ (computed similarly as $\hat{\theta}$ given above) when the ith observation is omitted from the original sample of size n, and denote the arithmetic mean of these estimates by

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}. \tag{14}$$

A quantity known as the ith "pseudo-value" is defined by

$$J_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}. \tag{15}$$

The jackknife estimator of θ is given by

$$J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} J_i = n\hat{\theta} - (n-1)\tilde{\theta}. \tag{16}$$

If the original estimate $\hat{\theta}$ is biased, then, under certain conditions, part of the bias is removed by the jackknife procedure, and an estimate of the standard error of the jackknife estimate, $J(\hat{\theta})$, is given by

$$\hat{\sigma}_{J(\hat{\theta})} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(J_i - J(\hat{\theta})\right)^2}. \tag{17}$$

Another application of the pseudo-values, suggested by J. Tukey (see Miller, 1974), is to use the pseudo-values to obtain confidence intervals for the parameter, θ, based on the following pivotal quantity:

$$t = \frac{J(\hat{\theta}) - \theta}{\hat{\sigma}_{J(\hat{\theta})}}. \tag{18}$$

The statistic, t, given by equation (18) has an approximate Student's t distribution with $n-1$ degrees of freedom, which can be used to derive the following approximately two-sided $(1-2\alpha)100\%$ confidence interval for θ:

$$\left(J(\hat{\theta}) - t_{\alpha, n-1}\hat{\sigma}_{J(\hat{\theta})},\ J(\hat{\theta}) + t_{\alpha, n-1}\hat{\sigma}_{J(\hat{\theta})}\right). \tag{19}$$

The upper limit of equation (19) is an approximate $(1-\alpha)100\%$ UCL for θ. If the sample size, n, is large, then the upper αth t-quantile can be replaced with the corresponding upper αth standard normal quantile, z_α. Observe also that when $\hat{\theta}$ is the sample mean, then the jackknife estimate is the sample mean, that is $J(\bar{X}) = \bar{X}$; the estimate of the standard error in equation (17) simplifies to $s_x/n^{1/2}$, and the confidence interval in equation (19) reduces to the familiar t-statistic based confidence interval given by equation (13).

Bootstrap Estimation

In the bootstrap procedure, repeated samples of size n are drawn with replacement from the given set of observations. The process is repeated a large number of times, and each time an estimate of θ is computed. The estimates thus obtained are used to compute an estimate of the standard error of $\hat{\theta}$. There exists in the literature of statistics an extensive array of different bootstrap methods for constructing confidence intervals. In this article two of these methods are considered: 1) the standard bootstrap, and 2) the pivotal (or Studentized) bootstrap method as discussed by Hall (1988). A general description of bootstrap methods, illustrated by application to the sample mean, follows:

Step 1. Let $(x_{i1}, x_{i2}, \ldots, x_{in})$ represent the ith sample of size n **with replacement** from the original data set (x_1, x_2, ..., $_{Xn}$). Then compute the sample mean and denote it by $\bar{X}_1$.

Step 2. Perform Step 1 independently N times (e.g., 500-1000), each time calculating a new estimate. Denote those estimates by $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_N$. The bootstrap estimate of the population mean is the arithmetic mean, $\bar{x}_B$, of the N estimates $\bar{x}_I$. The bootstrap estimate of the standard error is given by

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(\bar{x}_i - \bar{x}_B\right)^2}. \tag{20}$$

If some parameter, θ (say, a population median), other than the mean is of concern, with an associated estimate (e.g., the sample median), then the same steps previously described could be applied with the parameter and its estimate used in place of μ_1 and $\bar{x}$. Specifically, the estimate, $\hat{\theta}_I$, would be computed, instead of $\bar{x}_I$, for each of the N bootstrap samples. The general bootstrap estimate, denoted by $\bar{\theta}_B$, is the arithmetic mean of the N estimates. The difference, $\bar{\theta}_B - \hat{\theta}$, provides an estimate of the bias of the estimate, $\hat{\theta}$, and the bootstrap estimate of the standard error of $\hat{\theta}$ is given by

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(\hat{\theta}_i - \bar{\theta}_B\right)^2}. \tag{21}$$

The standard bootstrap confidence interval is derived from the following pivotal quantity:

$$z = \frac{\hat{\theta} - \theta}{\hat{\sigma}_B}. \tag{22}$$

Finally, the (1-2α)100% standard bootstrap confidence interval for θ, which assumes that equation (22) is approximately normal, is

$$\left(\hat{\theta} - z_{\alpha}\hat{\sigma}_{B},\ \hat{\theta} + z_{\alpha}\hat{\sigma}_{B}\right). \tag{23}$$

In this case, the bootstrap approach gives a convenient way to estimate the standard error of $\hat{\theta}$. Depending on the type of estimate $\hat{\theta}$, the standard error may be quite difficult to derive, and consequently difficult to estimate. However, the bootstrap approach always yields an estimate of the standard error directly from the data, even when the mathematical form of the standard error is not known.

Another variation of the bootstrap method, called the "bootstrap *t*" by Efron (1982), is a nonparametric procedure which uses the bootstrap methodology to estimate quantiles of the pivotal quantity in equation (12). As previously noted, for nonnormal populations the required quantiles may not be easily obtained, or it may be impossible to compute exactly. However, with a variation of the bootstrap procedure, as proposed by Hall (1988), the required quantiles can be estimated directly from the data. Specifically, in Steps 1 and 2 described above, if $\bar{x}$ is the sample mean computed from the original data, and $\bar{x}_i$ and $s_{x,\,i}$ are the sample mean and sample standard deviation computed from the *i*th resampling of the original data, the *N* quantities, $t_i = (\bar{X}_i - \bar{X})/s_{x,\,i}$, are computed and sorted, yielding ordered quantities $t_{(1)} \le t_{(2)} \le \cdots \le t_{(N)}$. The estimate of the upper αth quantile of the pivotal quantity in equation (12) is $t_{\alpha,\,B} = t_{((1-\alpha)N)}$. For example, if $N = 1000$ bootstrap samples are generated, then the 950th ordered value, $t(950)$, would be the bootstrap estimate of the upper .05th quantile of the pivotal quantity in equation (12). This estimated quantile can be used in place of the upper αth Student's *t* quantile in an interval of the form given in equation (13). In the next section, this method of construction will be called the "pivotal bootstrap". This approach has the advantage that it does not rely on the assumption of a special parametric form for the distribution of the population, and it does not require an assumption of approximate normality for the pivotal quantity as does the standard bootstrap interval of equation (23).

In the examples to follow, the jackknife, the standard bootstrap method, and the pivotal bootstrap methods are applied using the sample mean, $\bar{x}$, and also the estimate given by equation (8), which is the MVUE of the mean when the population is lognormal.

The Central Limit Theorem

Given a random sample, $x_1, x_2, \ldots, x_n$, of size *n* from a population with a finite variance, σ_1^2, where $\theta = \mu_1$ is the unknown population mean, the Central Limit Theorem (CLT) states that the asymptotic distribution, as *n* approaches infinity, of the sample mean, $\bar{x}_n$, is normally distributed with mean, μ_1, and variance, σ_1^2/n. More precisely, the sequence of random variables

$$z_n = \frac{\bar{x}_n - \mu_1}{\sigma_1/\sqrt{n}} \tag{24}$$

has a standard normal limiting distribution. In practice, this means that for large sample sizes *n*, the sample mean, $\bar{x}$, has an approximate normal distribution irrespective of the underlying distribution function. Consequently, equation (24) is an approximate pivotal quantity for large *n*. This powerful result can be used to obtain approximate (1-2α)100% confidence intervals for the mean for any distribution with a finite variance, although, strictly speaking, it requires one to know the population standard deviation, σ_1. However, as noted by Hogg and Craig (1978), if σ_1 is replaced by the sample standard deviation, s_x, the normal approximation for large *n* is still valid. This leads to the following confidence interval:

$$\left(\bar{x} - z_{\alpha}s_x/\sqrt{n},\ \bar{x} + z_{\alpha}s_x/\sqrt{n}\right). \tag{25}$$

Note that the confidence interval in equation (25) has the same general form as equation (13), but with the *t* quantiles replaced with approximate standard normal quantiles. As noted previously, if the lower limit is disregarded, the upper limit of the interval provides a one-sided UCL for the population mean.

An often cited rule of thumb for a sample size with the CLT is $n \ge 30$. However, this may not be adequate if the population is highly skewed. A refinement of the CLT approach, which makes an adjustment for skewness, is discussed by Chen (1995). Specifically, the "adjusted CLT" UCL is obtained if the standard normal quantile, z_α, in the upper limit of equation (25) is replaced by

$$z_{\alpha, adj} = z_\alpha + \frac{\hat{\kappa}_3}{6\sqrt{n}}\left(1 + 2z_\alpha^2\right) \tag{26}$$

where $\hat{\kappa}_3$ is the sample coefficient of skewness,

$$\hat{\kappa}_3 = \frac{1}{n s_x^3} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3 . \tag{27}$$

Notice that this adjustment results in a UCL which is larger than that of equation (25) when the sample skewness is positive.

The Chebychev Theorem

This theorem is given here to obtain a reasonably conservative estimate of the UCL of the mean. The two-sided Chebychev theorem states that given a random variable, X, with finite mean and standard deviation, μ_1 and σ_1, one has

$$\mathrm{P}\left(-k\sigma_1 \le X - \mu_1 \le k\sigma_1\right) \ge 1 - 1/k^2 . \tag{28}$$

This result can be applied with the sample mean, $\bar{x}$, to obtain a conservative UCL for the population mean. Specifically, if the right side of equation (28) is equated to 0.95, then $k = 4.47$, and $UCL = \bar{x} + 4.47\sigma_1/n^{1/2}$ is a conservative 95% upper confidence limit for the population mean. Of course, this would require the user to know the value of σ_1. The obvious modification would be to replace σ_1 with the sample standard deviation, s_x, but, since this is estimated from the data, the result is no longer guaranteed to be conservative. In general, if μ_1 is an unknown mean, $\hat{\mu}_1$ is an estimate and $\hat{\sigma}(\hat{\mu}_1)$ is an estimate of the standard error of $\hat{\mu}_1$, then the quantity $UCL = \hat{\mu}_1 + 4.47\hat{\sigma}(\hat{\mu}_1)$ will give 95% UCLs for μ_1, which should tend to be conservative, but this is not assured. This could be used, for example, with the mean of a lognormal population, using equation (8), as the estimate of the population mean and the square root of equation (10) as the estimate of the standard error. This has been used in the following examples.

4. Examples

Monte carlo simulation experiments were performed to compare various methods of computing the UCL of the lognormal mean. Based on these experiments, the methods of jackknife, bootstrap, or even the conservative method based on the Chebychev inequality appear to be superior to the H-statistic-based UCL for small sample sizes. When the number of samples is large ($n \ge 100$), all of these methods give similar results. In this section, a few simulated examples are provided to compare the various methods of computing values of the UCL. A few examples from Superfund sites have also been included.

Example 4.1. Simulated sample from a mixture of two normal populations, N(100, 50) and N(1000, 100).

This example uses the sample of size n = 15 which was discussed previously in Example 2.1. Recall, that this is a simulated sample from a mixture of two normal populations. The mean of the mixed normal population is μ_1 = 400. The values of the mean, standard deviation, and coefficient of variation computed for the log-transformed data are:

$\bar{y}$ *= 5.090,* s_y *= 1.705, and* CV_y *= 0.34.*

The values of the mean, standard deviation, and CV computed for the raw data are:

$\bar{x}$ *= 403.35,* s_x *= 453.94, and* CV_x *= 1.125.*

If it is assumed (incorrectly) that the population is lognormal, point estimates based on MVUE theory of the mean, μ_1, standard deviation, σ_1, and standard error of the mean are 572.98, 1334.56 and 290.14, respectively. Estimates of the 80th, 90th, and 95th percentiles of a lognormal distribution are 686.33, 1453.48, and 2685.56, respectively.

Discussion of Example 4.1

The 95% UCL values obtained from the methods discussed above, without using lognormal theory, are:

Jackknife	609.75
Standard Bootstrap	584.32
Pivotal Bootstrap	651.52
CLT	596.16
Adjusted CLT	618.51
Chebychev	927.27

The values of the 95% UCL obtained from the methods discussed above, calculated using the lognormal theory, are:

Jackknife	1085.17
Standard Bootstrap	994.40
Chebychev	1869.90
H-UCL	4150.96

Notice that the 95% UCL computed from the H-statistic (4150.96) exceeds the estimated 95th percentile (2685.56) of an assumed lognormal distribution. The H-UCL is also an order of magnitude larger than the true mean, 400, of the mixture of two normal populations.

It is also of interest to see how the methods compare when applied to simulated lognormal data with different sample sizes and various combinations of parameter values.

Example 4.2. Simulated sample of size n = 15 from a lognormal distribution, LN(5, 1).

In this example, n = 15 data were generated from the lognormal distribution LN(5,1), with following (true) values of population parameters: μ_1 = 244.69, σ_1 = 320.75, and CV = 1.31. The generated data are:

139.2056, 259.9746, 138.7997, 48.8109, 166.1733, 54.1241, 120.3665, 60.9887, 551.2073, 66.3336, 16.0695, 364.5569, 153.2404, 271.5436, 473.6461.

The values of the sample mean, standard deviation, and CV of the log-transformed data are:

$\bar{y}$ = 4.887, s_y = 0.966, CV_y = 0.20.

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x}$ = 192.34, s_x = 161.56, CV_x = 0.84.

For a lognormal distribution, the estimates of μ_1, σ_1, and the standard error of the mean, based on MVUE theory, are 202.58, 219.21, and 54.00, respectively. The MLEs of μ_1, σ_1, and CV are 211.33, 262.47, and 1.24, respectively. Estimates of the 80th, 90th, and 95th percentiles of the lognormal distribution are 299.79, 458.58, and 649.31, respectively.

Discussion of Example 4.2

The values of the 95% UCL obtained from the methods discussed above, without using lognormal theory, are:

Jackknife	265.79
Standard Bootstrap	258.21
Pivotal Bootstrap	292.17
CLT	260.96
Adjusted CLT	271.57
Chebychev	378.80

The values of the 95% UCL obtained from the methods discussed above, calculated from lognormal theory, are:

Jackknife	289.30
Standard Bootstrap	281.22
Chebychev	448.41
H-UCL	427.62

The differences in UCLs for the various methods are not as extreme as they were in the previous example, but a similar pattern with the Chebychev (as expected) and H-UCL limits being the largest is still present. However, unlike the previous example, the 95% UCL is below the estimated 95th percentile of a lognormal distribution, as one would intuitively expect. It is also interesting to note that the CV estimated as the ratio of the sample standard deviation to the sample mean from raw data is less than 1 (0.84), while the CV computed from the MLEs is slightly greater than 1 (1.24). According to the CV test, which says that if CV <1.0, then the population is normally distributed, the former CV of 0.84 might lead one to incorrectly assume that the population is normally distributed.

In the next example, the variance of the log-transformed variable is increased slightly with a corresponding increase in CV and skewness.

Example 4.3. Simulated sample of size n = 15 from a lognormal distribution, LN(5, 1.5).

In this example, n = 15 observations were generated from the lognormal distribution, LN(5,1.5), with the following true values of population parameters: μ_1 = 457.14, σ_1 = 1331.83, CV = 2.91. The generated data are:

440.8517, 1013.4986, 1857.7698, 500.9632, 397.9905, 110.7144, 196.2847, 128.2843, 1529.9753, 5.7978, 940.8903, 597.5925, 1519.5159, 181.6512, 52.8952.

The sample mean, standard deviation, and CV of the log-transformed data are:

$\bar{y}$ = 5.761, s_y = 1.536, and CV_y = 0.27.

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x}$ = 631.65, s_x = 603.13, and CV_x = 0.96.

For a lognormal distribution, the estimates of μ_1, σ_1, and standard error of the mean, based on MVUE theory, are 894.76, 1784.95, and 405.79, respectively. The MLEs of μ_1, σ_1, and CV are 1033.63, 3202.28 and 3.10, respectively. Estimates of the 80th, 90th, and 95th percentiles of the lognormal distribution are 1163.05, 2286.63, and 3975.71, respectively.

Discussion of Example 4.3

The values of the 95% UCL obtained from the methods discussed above, without using lognormal theory, are:

Jackknife	905.88
Standard Bootstrap	882.82
Pivotal Bootstrap	977.18
CLT	887.82
Adjusted CLT	919.81
Chebychev	1327.75

The values of the 95% UCL obtained from the methods discussed above, calculated from lognormal theory, are:

Jackknife	1534.94
Standard Bootstrap	1363.26
Chebychev	2708.63
H-UCL	4570.27

As in the case of Example 4.1, the 95% H-UCL (4570.27) again exceeds the estimated 95th percentile of the lognormal distribution. The situation with the CV is similar to that of Example 4.2. That is, the CV computed from raw data (0.96) is less than 1, which by application of the CV-test could lead one to adopt (incorrectly) the normal distribution. Notice that the true CV and the estimate based on the MLEs are both close to three. The next example involves the same population but with a larger sample size.

Example 4.4. Simulated sample of size n = 31 from a lognormal distribution, LN(5, 1.5).

In this simulated example, n = 31 observations were generated from a lognormal distribution, LN(5,1.5). This is the same distribution use in the previous example, and thus true mean, standard deviation, and CV are the same. The generated data are:

49.0524, 806.8449, 122.2339, 697.7315, 2888.1238, 37.7998, 7.2799, 292.5909, 433.4413, 639.7468, 3876.8206, 1376.8859, 197.8634, 93.0379, 180.9311, 1817.9912, 284.3526, 344.6761, 44.8680, 297.3899, 11.9195, 100.5519, 264.7574, 41.3961, 43.4202, 1053.3770, 2067.0361, 132.2938, 75.9661, 53.2236, 83.5585.

The sample mean, standard deviation, and CV of log-transformed data are:

$\bar{y}$ = 5.326, s_y =1.577, and CV_v = 0.30

The sample mean, standard deviation, and CV for raw data are:

$\bar{x}$ = 594.10, s_x = 919.05, and CV_x = 1.55.

For a lognormal distribution, the estimates of μ_1, σ_1, and the standard error of the mean are 657.45, 1632.25, and 238.86, respectively. The MLEs of μ_1, σ_1, and CV are 713.34, 2369.11, and 3.32. Estimates of the 80th, 90th, and 95th percentiles of a lognormal distribution are 779.73, 1560.71, and 2753.62, respectively.

Discussion of Example 4.4

The values of the 95% UCL obtained from the methods discussed above, without using lognormal theory, are:

Jackknife	874.22
Standard Bootstrap	854.51
Pivotal Bootstrap	1003.00
CLT	865.64
Adjusted CLT	932.36
Chebychev	1331.95

The values of the 95% UCL obtained from the methods discussed above, calculated from lognormal theory, are:

Jackknife	1062.35
Standard Bootstrap	1088.94
Chebychev	1725.15
H-UCL	1792.54

As one might expect with a larger sample size (n = 31), the point estimates tend to be closer to the true parameter values they are intended to estimate. Also, there is not as much variation among the UCLs computed from the different methods. Furthermore, the H-UCL is below the estimated 95th percentile of the lognormal distribution.

In the next example, a sample of size n = 15 is considered again, but with the variance of the log-transformed variable slightly larger than that of Examples 4.2-4.4.

Example 4.5. Simulated sample of size n = 15 from a lognormal distribution, LN(5, 1.7).

This last simulated data set of size n = 15 is obtained from LN(5, 1.7), with the following true values of population parameters: μ_1 = 629.55, σ_1 = 2595.18, CV = 4.12.

The generated data are:

16.5197, 235.4977, 1860.4443, 74.5825, 3.9684, 325.2712, 167.7949, 189.0130, 1307.6180, 878.8519, 35.4675, 96.2498, 229.2540, 182.0494, 1498.6146.

The sample mean, standard deviation, and CV of the log-transformed data are:

$\bar{y}$ = 5.178, s_y = 1.710, CV_y = 0.33.

The sample mean, standard deviation, and CV for raw data are:

$\bar{x}$ = 473.41, s_x = 606.79, CV_x = 1.28.

For a lognormal distribution, the estimates of μ_1, σ_1, and the standard error of the mean, based on MVUE theory, are 629.82, 1473.12, and 319.0, respectively. The MLEs of μ_1, σ_1, and CV are 765.52, 3213.52, and 4.20, respectively. Estimates of the 80th, 90th, and 95th percentiles for a lognormal distribution are 752.50, 1596.91, and 2955.58, respectively.

Discussion of Example 4.5

The values of the 95% UCL obtained from the four methods discussed above, without using lognormal theory, are:

Jackknife	749.31
Standard Bootstrap	721.07
Pivotal Bootstrap	862.51
CLT	731.14
Chebychev	1173.74

The values of the 95% UCL obtained from the four methods discussed above, calculated from lognormal theory, are:

Jackknife	1176.39
Standard Bootstrap	1141.95
Chebychev	2059.47
H-UCL	4613.32

Notice that in this example (as with Examples 4.1 and 4.3), the 95% H-UCL (4613.32) exceeds the estimated 95th percentile (2955.58) of the lognormal distribution.

The sample size and the mean of the log-transformed variable in examples 4.2, 4.3, and 4.5 are held constant at 15 and 5, respectively, whereas the standard deviation (sd) of the log-transformed variable are 1.0, 1.5, and 1.7, respectively. From these examples alone, it can be seen that as soon as the sd of the log-transformed variable becomes greater than 1.0, the H-statistic-based UCL becomes orders of magnitude higher than the largest concentrations observed, even when the data were obtained from a lognormal population. Thus, even though the H-UCL is theoretically sound and possesses optimal properties for truly lognormal populations such as being MVUE, the practical merit of the use of H-UCL in environmental applications is questionable when the sd of the log-transformed variable starts exceeding 1.0. This is especially true for small sample sizes (e.g., n <30). As seen in the examples discussed here, the use of the lognormal distribution and the H-UCL in some circumstances tends to hide contamination rather than find it, which is contrary to one of the main objectives in many environmental applications. Actually, under the assumption of lognormal distribution, one can get away with very little or no cleanup, (Bowers, Neil, and Murphy 1994), at a polluted site.

Example 4.6. Data from the Naval Construction Battalion Center (NCBC) Superfund Site in Rhode Island.

Inorganic analyses were performed on the groundwater samples from seventeen (17) wells from the NCBC Site. The main objective was to provide reliable estimates of the mean background threshold levels for the various inorganic contaminants at the site. The UCLs have been computed using the procedures described above. The results for two of the contaminants, aluminum and manganese, are summarized below.

Aluminum: *290, 113, 264, 2660, 586, 71, 527, 163, 107, 71, 5920, 979, 2640, 164, 3560, 13200, 125.*

The sample mean, standard deviation, and CV of log-transformed data are:

$\bar{y}$ = 6.226, s_y = 1.659, CV_y = 0.27.

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x}$ = 1849.41, s_x = 3351.27, CV_x = 1.81.

With the lognormal assumption, the estimates of μ_1, σ_1, *and the standard error of the mean, based on MVUE theory, for aluminum are 1704.84, 3959.87, and 807.64, respectively. The MLEs of* μ_1, σ_1, *and CV are 2002.71, 7676.37, and 3.83, respectively. Estimates of the 80th, 90th and 95th percentiles for a lognormal distribution are 2054.44, 4263.44, and 7747.81, respectively.*

Manganese*: 15.8, 28.2, 90.6, 1490, 85.6, 281, 4300, 199, 838, 777, 824, 1010, 1350, 390, 150, 3250, 259.*

The sample mean, standard deviation, and CV of log-transformed data are:

$\bar{y}$ = 5.91, s_y = 1.568, CV_y = 0.27.

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x}$ = 902.25, s_x = 1189.49, CV_x = 1.32.

With the lognormal assumption, the estimates of μ_1, σ_1, *and the standard error of the mean, based on MVUE theory, for manganese are 1100.92, 2340.72, and 490.16, respectively. The MLEs of* μ_1, σ_1, *and CV are 1262.59, 4125.5, and 3.27, respectively. Estimates of the 80th, 90th, and 95th percentiles for a lognormal distribution are 1389.65, 2769.95, and 4870.45, respectively.*

The calculated Shapiro Wilks statistics for the raw data are 0.594 (aluminum) and 0.725 (manganese), and for the log-transformed data, the corresponding values are 0.913 and 0.969. The tabulated critical value for 0.10 level of significance is 0.91. Thus, for both aluminum and manganese, the data failed the normality test and passed the lognormality test at significance level 0.10 (Note: Shapiro-Wilks is a lower tail test).

Discussion of Example 4.6

The values of the 95% UCL obtained from the methods discussed above, without using lognormal theory, are:

	Aluminum	Manganese
Jackknife	3268.22	1405.83
Standard Bootstrap	3125.56	1354.15
Pivotal Bootstrap	5286.63	1968.03
CLT	3186.47	1376.82
Adjusted CLT	3675.94	1503.84
Chebychev	5482.64	2191.81

Observe that for both of the contaminants, the 95% UCLs calculated from the jackknife, both bootstrap methods, the CLT, the adjusted CLT, and the Chebychev limit are well below their respective estimates of the 95th percentile (Aluminum: 7747.81 and Manganese: 4870.45) of assumed (based on Shapiro-Wilks' test) lognormal distributions.

The values of the 95% UCL obtained from the methods discussed above, calculated from lognormal theory, are:

	Aluminum	Manganese
Jackknife	3283.34	1889.52
Standard Bootstrap	3663.20	1821.55
Chebychev	5314.99	3291.95
H-UCL	9102.73	5176.16

Observe that the 95% UCLs calculated using lognormal theory from the jackknife, the bootstrap, and the Chebychev inequality are similar to the respective values obtained without using lognormal theory, and that these are well below their respective estimated 95th percentiles for a lognormal distribution. The 95% UCLs calculated from the H-statistic, however, exceed their respective estimated 95th percentiles for a lognormal distribution.

Example 4.7. Data from the Elrama School Superfund site in Washington County, PA.

The data were compiled from two waste piles for risk evaluations of the contaminants found at the Elrama School Superfund Site, Washington County, PA. Twenty-six (26) contaminants (10 inorganics, 12 semi-volatile compounds, and 4 volatile compounds) were detected in both of the waste piles. Using the nonparametric Kolmogorov-Smirnov two-sample test on the two waste piles, it was concluded that there is no statistically significant difference between distributions of the contaminants from the two waste piles. Thus, the data from these two waste piles were combined to compute all of the relevant statistics such as the mean, the standard deviation, and the UCLs. This resulted in data sets consisting of 23 observations (15 from Waste Pile 1 and 8 from Waste Pile 2). The results are provided for two of the contaminants of concern: aluminum and toluene.

***Aluminum**: 31900.0, 8030.0, 12200.0, 11300.0, 4770.0, 5730.0, 5410.0, 8420.0, 8200.0, 9010.0, 8600.0, 9490.0, 9530.0, 7460.0, 7700.0, 13700.0, 30100.0, 7030.0, 2730.0, 5820.0, 8780.0, 360.0, 7050.0.*

The sample mean, standard deviation, and CV of the log-transform data are:

$\bar{y}$ = 8.927, s_y = 0.845, CV_y = 0.095

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x}$ = 9709.57, s_x = 7310.02, CV_x = 0.75.

With the lognormal assumption, the estimates of μ_1, σ_1, and the standard error of the mean, based on MVUE theory, for aluminum are 10552.68, 10031.60, and 2044.90, respectively. The MLEs of μ_1, σ_1, and CV are 10768.22, 10993.32, and 1.02, respectively. Estimates of the 80th, 90th, and 95th percentiles for a lognormal distribution are 15323.48, 22224.45, and 30381.95, respectively.

***Toluene**: 7300.0, 6.0, 6.0, 5.5, 29000.0, 46000.0, 12000.0, 2500.0, 1300.0, 3.0, 510.0, 230.0, 63.0, 6.0, 5.5, 6.0, 6.0, 5.5, 280000.0, 8.0, 28.0, 6.0, 7.0.*

The sample mean, standard deviation and CV of log-transform data are:

$\bar{y}$ = 4.652, s_y = 3.660, CV_y = 0.79

The sample mean, standard deviation, and CV for the raw data are:

$\bar{x} = 16478.33$, $s_x = 58510.78$, $CV_x = 3.55$.

With the lognormal assumption, the estimates of μ_1, σ_1, and the standard error of the mean, based on MVUE theory, for toluene are 21328.39, 362471.55, and 18788.05, respectively. The MLEs of μ_1, σ_1, and CV are 84702.17, 68530556.56, and 809.08, respectively. Estimates of the 80th, 90th, and 95th percentiles for a lognormal distribution are 2264.17, 11329.16, and 43876.88, respectively.

The Shapiro-Wilks statistics for the raw data are 0.707 (aluminum) and 0.313 (toluene), and for the log-transformed data, the corresponding values are 0.781 and 0.818. The tabulated critical value for a 0.10 level of significance with n = 23 is 0.928. Thus, neither a normal nor a lognormal distribution gives a good fit.

Discussion of Example 4.7

The values of the 95% UCL obtained from the methods discussed above, without using lognormal theory, are:

	Aluminum	Toluene
Jackknife	12327.40	37431.95
Standard Bootstrap	12246.67	33494.25
Pivotal Bootstrap	15161.90	152221.00
CLT	12216.95	36547.89
Adjusted CLT	12895.10	47316.80
Chebychev	16522.94	71013.85

The values of the 95% UCL obtained from the four methods discussed above, calculated from lognormal theory, are:

	Aluminum	Toluene
Jackknife	13542.11	62263.37
Standard Bootstrap	13579.18	278888.51
Chebychev	19693.40	105757.50
H-UCL	16503.51	18444955.15

Observe that the 95% UCL for toluene, calculated from the H-statistic, is orders of magnitude higher than those calculated from the other methods, and is also orders of magnitude higher than the maximum observed toluene concentration at the site. Also, with the toluene data, the pivotal bootstrap method results in a UCL which is two to five times larger than the others computed from the non-lognormal theory methods. It is even larger than the Chebychev limit. As noted earlier, this is possible when the standard error of the point estimate is also estimated from the data. In most environmental applications, the true population standard deviation of the point estimate is unknown, and therefore, it needs to be estimated from the available data. Note, however, it is two orders of magnitude smaller than the H-UCL.

Note, also, that the CV (0.75) computed from the raw data for aluminum is less than 1. The use of the CV-test for normality could lead one to assume normality, even though the Shapiro-Wilks test strongly rejects the normal distribution (p-value = 0.00002).

5. Summary and Recommendations

It is seen from the simulated examples that, even when the underlying distribution is lognormal, the performance (in terms of a lower UCL) of the jackknife, bootstrap, and CLT procedures is more accurate than that of the H-UCL. In each of the four simulation experiments, the 95% UCLs computed from all of the above methods exceeds the true respective population means, but the 95% H-UCL is consistently larger, except in some cases where it is comparable to the conservative Chebychev result, than the 95% UCLs obtained from other methods. It is also seen from the simulation examples that the estimate of the CV based on the MLEs is closer to the true CV than the usual (moment) estimate of CV. Furthermore, the usual estimate of the CV appears to underestimate the true CV. In some of the examples, the usual estimate of the CV is less than 1, while the true population CV is somewhat greater than 1. That is, the rule of thumb (CV-test) which declares the distribution to be normal when the moment estimate of the CV is less than 1, can frequently lead to an incorrect assumption about the underlying distribution of the data.

Moreover, from the examples discussed in this paper, it is observed that the H-UCL becomes order of magnitudes higher even when the data were obtained from a lognormal population and can lead to incorrect conclusions. This is especially true for samples of smaller sizes (e.g., <30). It appears that the lognormal distribution and the H-UCL tend to hide contamination rather than revealing it. Under the assumption of the lognormal distribution, one can get away with very little or no cleanup at a polluted site. Thus, although the H-UCL is theoretically sound and possesses optimal properties, the practical merit of the H-UCL in environmental applications is questionable, as it becomes an order of magnitude higher than the largest observed concentration when the sd of the log-transformed data starts exceeding 1.0. It is therefore, recommended that in environmental applications, the use of the H-UCL to obtain an estimate of the upper confidence limit of the mean should be avoided.

Based on the monte carlo simulation results, and the authors' experience with Superfund site work, the following steps for computing a UCL of the mean of the contaminant(s) of concern are recommended:

1) Plot histograms of the observed contaminant concentrations and perform a statistical test of normal or lognormal distribution (e.g., the Shapiro-Wilks test). *Do not use the rule of thumb that declares the data distribution to be normal if CV is less than 1.*

2) If a normal distribution provides an adequate fit to the data, then use the Student's *t* approach (equivalent to the jackknife) for calculating the UCL of the population mean.

3) If a lognormal distribution provides an adequate fit to the data, then a) use the lognormal theory based formulas for computing the MVUE of the population mean and the standard deviation, b) either use these MVUEs with the jackknife or bootstrap methods to calculate a UCL of the mean, or use the Chebychev approach for calculating a UCL. *Do not use the UCL based on the H-statistic, especially if the number of samples is less than 30.*

4) If the data distribution turns out to be neither normal nor lognormal, then use the nonparametric versions of the jackknife or bootstrap to calculate a UCL. Even if the lognormal distribution seems to provide a reasonable fit to the data, and if there is evidence of a mixture of two or more subpopulations, or if outliers are suspected, then using one of the nonparametric methods discussed above is recommended.

Notice

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and prepared this Issue Paper. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

References

Aitchison, J., and Brown, J. A. C. (1976), *The Lognormal Distribution*, Cambridge: Cambridge University Press.

Bain, L. J. and Engelhardt, M. (1992), *Introduction to Probability and Mathematical Statistics*, Boston: Duxbury Press.

Bowers, T., Neil, S., and Murphy, B. (1994), "Applying Hazardous Waste Site Cleanup Levels: A Statistical Approach to Meeting Site Cleanup Goals on Average." *Environmental Science and Technology.*

Bradu, D., and Mundlak, Y. (1970), "Estimation in Lognormal Linear Models," *Journal of the American Statistical Association*, 65, 198-211.

Chen, L. (1995), "Testing the Mean of Skewed Distributions," *Journal of the American Statistical Association*, 90, 767-772.

Efron, B. (1981), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Resampling Plans," *Biometrika*.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.

Efron, B., and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 36-48.

EPA (1992), "Supplemental Guidance to RAGS: Calculating the Concentration Term," Publication 9285.7-081, May 1992.

Finney, D. J. (1941), "On the Distribution of a Variate Whose Logarithm is Normally Distributed," *Journal of the Royal Statistical Society*, 7, 155-161.

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

Gilbert, R.O. (1993), "Comparing Statistical Tests for Detecting Soil Contamination Greater that Background," Pacific Northwest Laboratory, Technical Report No. DE 94-005498.

Hall, P. (1988). "Theoretical comparison of bootstrap confidence intervals," Ann. Statist., 16, 927-953.

Hogg, R.V., and Craig, A.T. (1978), *Introduction to Mathematical Statistics*, New York: Macmillan Publishing Company.

Land, C. E. (1971), "Confidence Intervals for Linear Functions of the Normal Mean and Variance," *Annals of Mathematical Statistics*, 42, 1187-1205.

Land, C. E. (1975), "Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance," in *Selected Tables in Mathematical Statistics*, vol. III, American Mathematical Society, Providence, R.I., 385-419.

Lechner, J.A. (1991), "Estimators for Type-II Censored (Log) Normal Samples," *IEEE Transactions on Reliability*, 40, 547-552.

Miller, R. (1974), "The Jackknife - A Review," *Biometrika*, 61, 1-15.

Power, M. (1992), "Lognormality in the Observed Size Distribution of Oil and Gas Pools as a Consequence of Sampling Bias," *Mathematical Geology*, 24, 929-945.

Singh, A. and Nocerino J.M. (1995), "Robust Procedures for the Identification of Multiple Outliers", *Chemometrics in Environmental Chemistry*, Statistical Methods, Vol 2., part G, 229-277, Springer Verlag, Germany.

Staudte, R. G., and Sheather, S. J. (1990), *Robust Estimation and Testing*, New York: John Wiley & Sons.

Stewart, S. (1994), "Use of Lognormal Transformations in Environmental Statistics," M.S. Thesis, Department of Mathematics, University of Nevada, Las Vegas.

Ott, W. (1990), "A Physical Explanation of the Lognormality of Pollutant Concentrations," *Journal of Air Waste Management Assoc.*, 40, 1378-1383.