United States Environmental Protection Agency | Office of Research and Development | Office of Solid Waste and Emergency Response | EPA/600/R-02/084 October 2002

Technology Support Center Issue

Estimation of the Exposure Point Concentration Term Using a Gamma Distribution

Anita Singh[1], Ashok K. Singh[2], and Ross J. Iaci[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.

This paper was prepared by Anita Singh, Ashok K. Singh, and Ross J. Iaci. 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 Christopher Sibert, Technology Support Center Director, at (702) 798-2270, Anita Singh at (702) 897-3234, or A. K. Singh at (702) 895-0364.

In Superfund and RCRA projects of the U.S. EPA, cleanup, exposure, and risk assessment decisions are often made based upon the mean concentrations of the contaminants of potential concern. A 95% upper confidence limit (UCL) of the population mean is used to estimate the exposure point concentration (EPC) term (EPA, 1992), to determine the attainment of cleanup standards (EPA, 1989), to estimate background level contaminant concentrations, or to compare the soil concentrations with the site specific soil screening levels (EPA, 1996). It is, therefore, important to compute an accurate and stable 95% UCL of the population mean from the available data.

The formula for computing a UCL depends upon the data distribution. Typically, environmental data are positively skewed, and

[1] Lockheed Martin Environmental Services, 1050 East Flamingo, Ste. E120, Las Vegas, NV 89119
[2] Department of Mathematics, University of Nevada, Las Vegas, NV 89154
[3] Department of Statistics, University of Georgia, Athens, GA 30602-1952

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

289CMB02.RPT ❖ Rev. 3/6/03

a lognormal distribution (EPA, 1992) is often used to model such data distributions. The H-statistic (Land, 1971) based upper confidence limit of the mean (denoted henceforth as H-UCL) is used in these applications. However, recent research in this area (Hardin and Gilbert, 1992; Singh, et al., 1997, 1999; and Schultz and Griffin, 1999) suggest that this may not be an appropriate choice. It is observed that for large values of standard deviation (e.g., exceeding 1.5 - 2.0) of the log-transformed data, the use of H-statistic leads to unreasonably large, unstable, and impractical UCL values. This is especially true for sample sets of smaller sizes (e.g., n < 20-25). The H-UCL is also very sensitive to a few low or high values. For example, the addition of a sample below detection limit can cause the H-UCL to increase by a large amount. Realizing that the use of H-statistic can result in an unreasonably large UCL, it has been recommended (EPA, 1992) to use the maximum observed value as an estimate of the UCL (EPC term) in cases where the H-UCL exceeds the maximum observed value. Also, when the sample size is 5 or less, the maximum observed concentration is often used as an estimate of the EPC term. However, it is observed that for highly skewed data sets, use of the maximum observed concentration may not provide the specified 95% coverage to the population mean (as shown in Section 5). This is especially true for samples of small size (e.g., 5-10). For larger sample sets/data sets (e.g., n$20), the use of the maximum observed value results in an overestimate of the 95% UCL of population mean. For such highly skewed data sets, use of a gamma distribution based UCL of the mean provides a viable option.

A positively skewed data set can quite often be modeled by lognormal as well as gamma distributions. Due to the relative computational ease, however, the lognormal distribution is used to model positively skewed data sets. However, use of a lognormal model for an environmental data set unjustifiably elevates the minimum variance unbiased estimate of the mean and its UCL to levels that may not be applicable in practice. In this paper, we propose the use of a gamma distribution to model positively skewed data sets. The objective of the present work is to study procedures which can be used to compute a stable and accurate UCL of the mean based upon a gamma distribution. Several parametric and non-parametric (e.g., standard bootstrap, bootstrap-t, Hall's bootstrap, Chebyshev inequality) methods of computing a UCL of the unknown population mean, μ, have also been considered. Monte Carlo simulation experiments have been performed to compare the performances of these methods. The comparison of the various methods has been evaluated in terms of the coverage (confidence coefficient) probabilities achieved by the various UCLs. Based upon this study, in Section 6, recommendations have been made about the computation of a UCL of the mean for skewed data distributions originating from various environmental applications.

1. Introduction

Suppose the Regional Project Manager (RPM) of a Superfund site believes that the mean concentration of the contaminant of potential concern (COPC) exceeds a specified cleanup standard, C_s, but the potentially responsible party (PRP) claims that the mean concentration is below C_s. In statistical terminology this can be stated in terms of testing of hypotheses. The hypotheses of interest are the null hypothesis that the mean concentration exceeds the cleanup standard, H_0: $\mu \geq C_s$, versus the alternative hypothesis, H_a: $\mu < C_s$. This formulation of the problem is protective of the environment because it assumes that the area in question is contaminated, and the burden of testing is to show otherwise. In order to perform a test of these hypotheses, a random sample is collected from the site and concentrations of the COPC in these samples are determined. A suitable statistical test is then used to make a decision.

A convenient way to perform a test of hypotheses about an unknown population parameter is first to compute a confidence interval for the parameter, and then reject H_0 if the hypothesized value, in this case the cleanup standard, C_s, lies outside of the confidence interval. For the one-sided hypotheses mentioned above, the test is based on a one-sided UCL of the mean. A one-sided UCL is a statistic such that the true population mean, μ, is less than the UCL with a prescribed probability or level of confidence, say

(1 ! "). For example, if the UCL is a 95% one-sided upper confidence limit, then μ < UCL with 95% confidence (or with 0.95 probability), and the set of all real numbers less than UCL forms a 95% upper one-sided confidence interval. The corresponding statistical test will reject H_0 (i.e., declare the site clean) if UCL < C_s, and the significance level of this test, or false positive error rate, is ". This follows because if the site is contaminated (i.e., μ $ C_s), then the probability of declaring it clean is the probability that UCL < C_s, which is at most ".

Testing of these hypotheses and computation of a UCL of the mean depends upon the population distribution of the COPC concentrations. Several procedures are available to compute a UCL of the mean of a normal or a lognormal distribution in the literature of environmental statistics (i.e., Singh, Singh, and Engelhardt, 1997, 1999; Schultz and Griffin, 1999). In this paper, we focus our effort on the inference procedures for an unknown population mean based upon a gamma distribution. The objective here is to study procedures that can be used to compute an accurate and stable UCL of the mean. Several parametric (Johnson, 1978; Chen, 1995; and, Grice and Bain, 1980) and non-parametric (e.g., standard bootstrap, bootstrap-t (Efron, 1982, Hall, 1988), Hall's bootstrap (Hall, 1992), Chebyshev inequality) methods of computing a UCL of population mean, μ, of a skewed distribution have also been considered. The comparison of the various methods has been performed in terms of the coverage (confidence coefficient) probabilities provided by the various 95% UCLs. Monte Carlo simulation experiments have been performed to compare the performances of these methods. Based upon this study, recommendations have been made about the computation of a UCL of the mean for skewed data distributions originating from various environmental applications.

Section 2 has a brief description of the gamma distribution and a discussion of goodness-of-fit tests for the gamma distribution. Section 2 also describes estimation of gamma parameters and the computation of the UCL of mean based upon a gamma distribution. Section 3 describes the various other methods which can be used to compute a UCL of population mean. Section 4 has some examples illustrating the procedures used. Section 5 discusses the Monte Carlo experiments used to illustrate these methods and results. Section 6 consists of our recommendations for dealing with heavily skewed data sets.

2. The Gamma Distribution

A continuous random variable, X (e.g., COPC concentration), is said to follow a two-parameter gamma distribution, G(k,2) with parameters k>0 and 2>0, if its probability density function is given by the following equation:

$$f(x; k, \theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}; x > 0 \qquad \textbf{(1)}$$

and zero otherwise. The parameter k is the shape parameter, and 2 is the scale parameter (the location parameter is set to zero). Plots of the gamma distribution, G(k,2) for varying choices of the shape parameter, k, and the scale parameter, 2, are shown in Figures 1-4. These figures have been generated using the statistical software package, MINITAB. The mean, variance, and skewness of a gamma distribution, G(k,2) are given as follows:

$$\text{Mean} = \mu = k\theta. \qquad \textbf{(2)}$$

$$\text{Variance} = \sigma^2 = k\theta^2. \qquad \textbf{(3)}$$

$$\text{Skewness} = 2/\sqrt{k}. \qquad \textbf{(4)}$$

From equation (4), it is noted that skewness increases as the shape parameter k decreases. Figures 1 and 2 have the graphs of highly skewed distributions. As k increases, skewness decreases, and consequently a gamma distribution starts approaching a normal distribution for larger values of k (e.g., k $10), as can be seen in Figures 3 and 4. Thus for larger values of k, the UCL based upon a gamma distribution and a UCL based upon a normal distribution are in close agreement. From Figures 1-4, it can also be seen that the scale parameter, 2, simply affects the scale of the distribution and has no effect on the shape of the gamma distribution. In practice, a highly skewed data set can be fitted by both lognormal and gamma distributions. However, the difference between the UCLs obtained using the two

distributions can be enormous. This is especially true when the shape parameter is small (e.g., $k < 1$). This is illustrated in examples 2-5 given in Section 4.

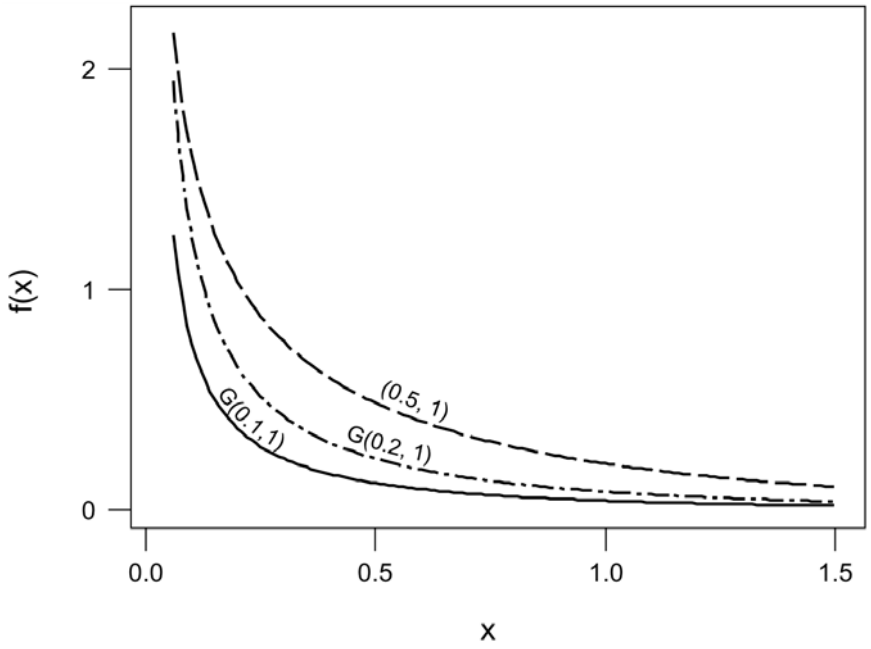

Figure 1. Graphs of the gamma distributions G(0.1, 1), G(0.2, 1), and G(0.5, 1).

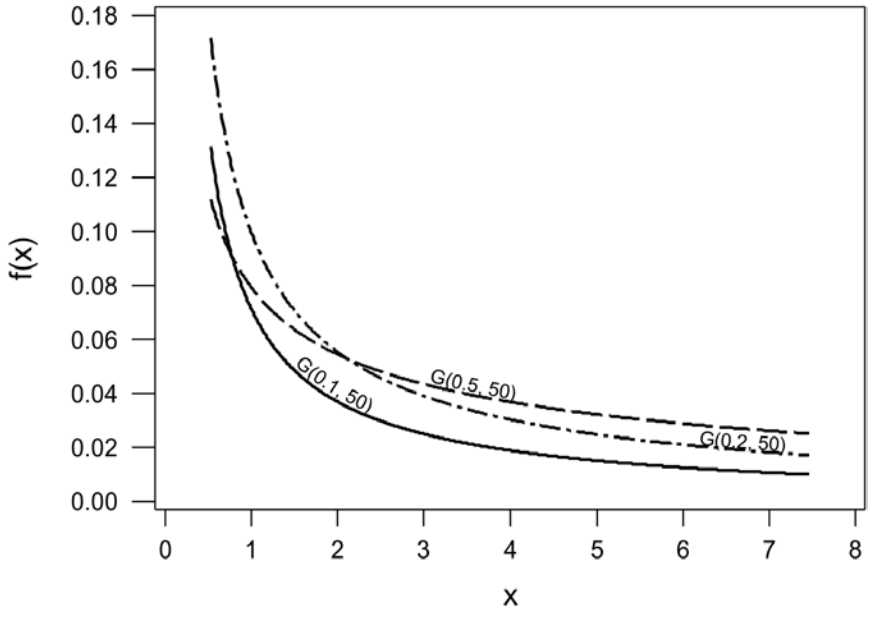

Figure 2. Graphs of the gamma distributions G(0.1, 50), G(0.2, 50), and G(0.5, 50).

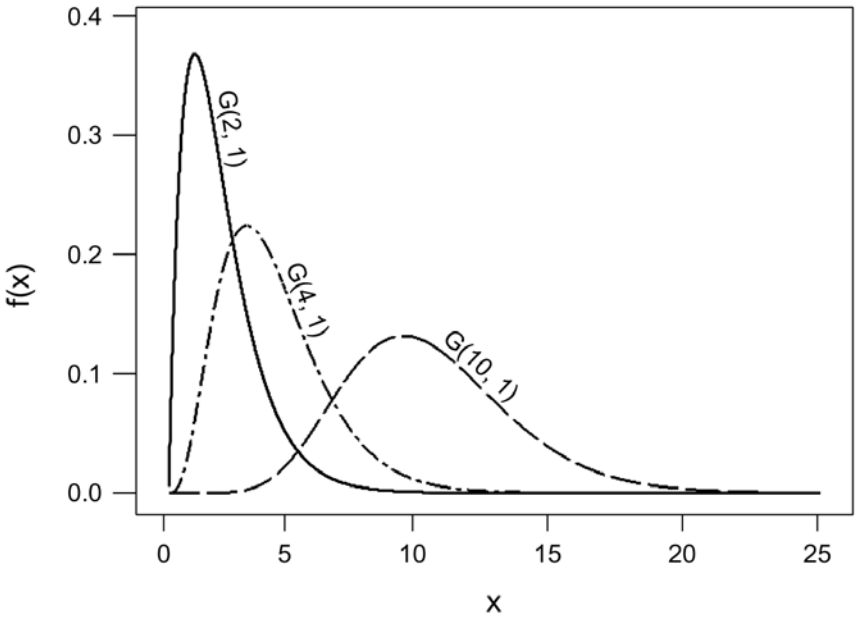

Figure 3. Graphs of the gamma distributions G(2, 1), G(4, 1), and G(10, 1).

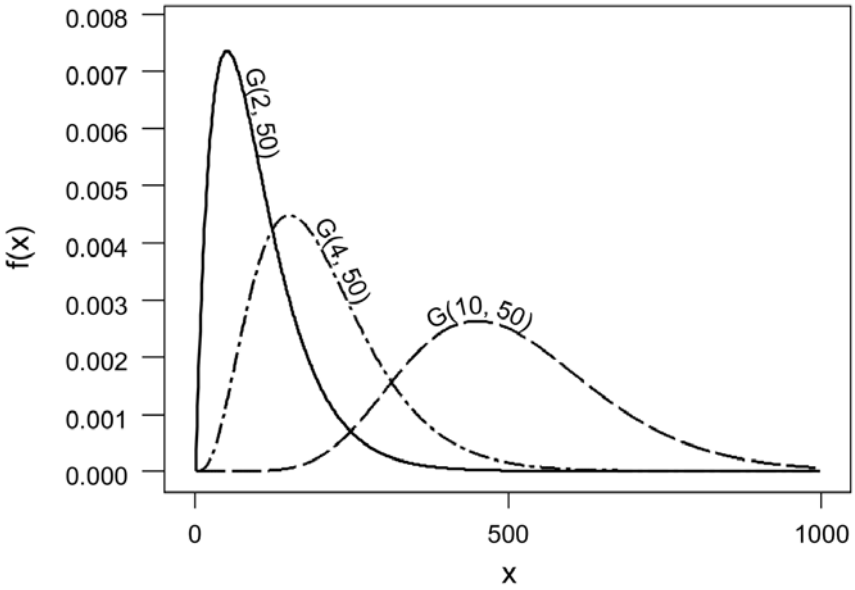

Figure 4. Graphs of the gamma distributions G(2, 50), G(4, 50), and G(10, 50).

2.1 Goodness-of-Fit Tests for Gamma Distribution

Since the goodness-of-fit tests for gamma distributions are not readily available, a brief description of those tests is given here. Several tests based upon empirical distribution functions (EDF) exist in the statistical literature, and can be used to test for a gamma distribution. These tests include Kolmogorov-Smirnov, D-test statistic, Anderson-Darling, A^2-test statistic, and Cramer-von Mises test statistics, W^2 and U^2 (e.g., see D'Agostino and Stephens (1986), page 101). The exact critical values of these statistics are not available; this is especially true when the shape

parameter, k, of the gamma distribution is less than 1. Some asymptotic upper-tail critical values of the test statistics, W^2, A^2, and U^2 are given in D'Agostino and Stephens (1986) for values of the shape parameter, k$1 (pages 152-155). Schneider (1978) also studied the goodness-of-fit tests for gamma distribution. He derived the critical values of Kolmogorov-Smirnov, D-test statistic for selected values of the shape parameter, k, and the sample size for the gamma distribution with unknown parameters. All of these tests are right-tailed. This means that if a computed test-statistic exceeds its respective "100% critical value, the null hypothesis of gamma distribution will be rejected at " level of significance.

Most of the commercially available software packages such as SAS and S-PLUS do not provide the goodness-of-fit tests for a gamma distribution. The software ExpertFit (developed by Law & Associates, Inc., 2001) performs a goodness-of-fit test for gamma distribution using the Anderson-Darling test statistic, A^2 and Kolmogorov-Smirnov test statistic. Due to the unavailability of the exact critical values of the general test statistics, the software ExpertFit (Law and Kelton (2000)) uses approximate critical values of the test statistic under the assumption that all parameters (e.g., shape and scale) of the distribution are known, that is the distribution is completely specified as given in Stephens (1970). Those critical values are the generic critical values for all completely specified distributions. ExpertFit uses these generic critical values to test for a gamma distribution. These critical values are also given on page 105 of D'Agostino and Stephens (1986). The authors of this article also developed a program, GamGood (2002), to test for a gamma distribution. This program computes the various goodness-of-fit test statistics using the formulae as given on page 101 of D'Agostino and Stephens (1986). In this paper, we also use the smoothed percentage points of the Kolmogorov-Smirnov (K-S), D-test statistic as computed by Schneider and Clickner (1976), Schneider (1978) to test for a gamma distribution. An illustration of the goodness-of-fit test for a gamma distribution has been discussed in Example 1.

Example 1

The following data set of size 20 is given by Grice and Bain (1980): 152, 152, 115, 109, 137, 88, 94, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, and 69. None of the parameters of the underlying distribution are known. The various goodness-of-fit test statistics are given by A^2 = 0.41496, W^2 = 0.06142, U^2 = 0.05111, and D = 0.13867. The estimated shape parameter, k, for this data set is 7.513 (see Example 1, to be continued). For a shape parameter of 7.513, the asymptotic 5% critical values (Table 4-21, page 155, D'Agostino and Stephens, 1986) of these statistics are: A^2 = 0.755, W^2 = 0.127, and U^2 = 0.117, and the critical value of the K-S statistic, is D = 0.196 (Table 7 of Schneider, 1978). Since all of the test-statistics are less than their respective critical values, it is concluded that there is insufficient evidence to conclude at the 0.05 level of significance that the data do not follow a gamma distribution.

2.2 Estimation of Parameters of the Gamma Distribution

Next, we consider the estimation of the parameters of a gamma distribution. The population mean and variance of a gamma distribution, G(k,2), are functions of both parameters, k and 2. In order to estimate the mean, one has to obtain estimates of k and 2. Computation of the maximum likelihood estimate (MLE) of k is quite complex and requires the computation of Digamma and Trigamma functions (Choi and Wette, 1969). Several authors (Choi and Wette, 1969, Bowman and Shenton, 1988, Johnson, Kotz, and Balakrishnan, 1994) have studied the estimation of shape and scale parameters of a gamma distribution. The maximum likelihood estimation procedure to estimate shape and scale parameters of a gamma distribution is described below.

Let $x_1, x_2, \ldots, x_n$ be a random sample (of COPC concentrations) of size n from a gamma distribution, G(k,2), with unknown shape and scale parameters k and 2, respectively. The log likelihood function is given as follows:

$$\log L(x_1, x_2, \ldots, x_n; k, \theta) = -nk \log(\theta) - n\log\Gamma(k) + (k-1)\sum \log x_i - \frac{1}{\theta}\sum x_i. \quad \textbf{(5)}$$

To find the MLEs of k and θ, which are $\hat{k}$ and $\hat{\theta}$, respectively, we differentiate the log likelihood function as given in (5) with respect to k and θ, and set the derivatives to zero. This results in the following two equations:

$$\log(\hat{\theta}) + \frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})} = \frac{1}{n}\sum \log(x_i), \text{ and} \qquad (6)$$

$$\hat{k}\hat{\theta} = \frac{1}{n}\sum x_i = \bar{x}. \qquad (7)$$

Solving equation (7) for $\hat{\theta}$ and substituting the result in equation (6), we get the following equation:

$$\frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})} - \log(\hat{k}) = \frac{1}{n}\sum \log(x_i) - \log\left(\frac{1}{n}\sum x_i\right). \qquad (8)$$

There does not exist a closed form solution of equation (8). This equation needs to be solved numerically for $\hat{k}$, which requires the computation of digamma and trigamma functions. This is quite easy to do using a personal computer. An estimate of k can be computed iteratively by using the Newton-Raphson (Faires and Burden, 1993) method leading to the following iterative equation:

$$\hat{k}_l = \hat{k}_{l-1} - \frac{\log(\hat{k}_{l-1}) - \Psi(\hat{k}_{l-1}) - M}{1/\hat{k}_{l-1} - \Psi'(\hat{k}_{l-1})}. \qquad (9)$$

The iterative process stops when $\hat{k}$ starts to converge. In practice, convergence is typically achieved in fewer than 10 iterations. In equation (9):

$$M = \log(\bar{x}) - \frac{1}{n}\sum \log(x_i),$$

$$\Psi(k) = \frac{d}{dk}\left(\log\Gamma(k)\right), \text{ and } \Psi'(k) = \frac{d}{dk}\left(\Psi(k)\right).$$

where Q (k) is the digamma function and QN (k) is the trigamma function. In order to obtain the MLEs of k and 2, one needs to compute the digamma and trigamma functions. Good approximate values for these two functions (Choi and Wette, 1969) can be obtained using the following approximations.

For $k$$8, these functions are approximated by:

$$\Psi(k) \approx \log(k) - \left\{1 + \left[1 - (1/10 - 1/(21k^2))/k^2\right]/(6k)\right\}/(2k), \qquad (10)$$

and

$$\Psi'(k) \approx \left\{1 + \left\{1 + \left[1 - (1/5 - 1/(7k^2))/k^2\right]/(3k)\right\}/(2k)\right\}/k. \qquad (11)$$

For $k < 8$, one can use the following recurrence relation to compute these functions:

$$\Psi(k) = \Psi(k+1) - 1/k, \qquad (12)$$

$$\text{and } \Psi'(k) = \Psi'(k+1) + 1/k^2. \qquad (13)$$

The iterative process requires an initial estimate of k. A good starting value for k in this iterative process is given by $k_0 = 1/(2M)$. Thom (1968) suggests the following approximation as an estimate of k:

$$\hat{k} \approx \frac{1}{4M}\left(1 + \sqrt{1 + \frac{4}{3}M}\right). \qquad (14)$$

Bowman and Shenton (1988) suggested using $\hat{k}$ as given by equation (14) to be a starting value of k for an iterative procedure, calculating $\hat{k}_l$ at the l^{th} iteration from the following formula:

$$\hat{k}_l = \frac{\hat{k}_{l-1}\left\{\log(\hat{k}_{l-1}) - \Psi(\hat{k}_{l-1})\right\}}{M}. \qquad (15)$$

Both equations (9) and (15) have been used to compute the MLE of k. It is observed that the estimate, $\hat{k}$ based upon Newton-Raphson method as given by equation (9) is in close agreement with that obtained using equation (15) with Thom's approximation as an initial estimate. Choi and Wette (1969) further concluded that the MLE of k, $\hat{k}$, is biased high. A bias corrected (Johnson, et al., 1994) estimate of k is given by the following equation:

$$\hat{k}^* = (n-3)\hat{k}/n + 2/(3n).16 \qquad (16)$$

In (16), $\hat{k}$ is the MLE of k obtained using either (9) or (15). Substitution of equation (16) in

equation (7) yields an estimate of the scale parameter, 2 given as follows:

$$\hat{\theta}^{*} = \bar{x} / \hat{k}^{*}. \qquad \textbf{(17)}$$

Next we provide an example illustrating the computations of the MLEs of k and 2.

Consider the data set of Example 1. The sample mean, $\bar{x}$, is 113.5. The MLEs of the two parameters, *k* and 2, are obtained iteratively using the Newton-Raphson method (equation 9), and Bowman and Shenton's proposal as given by equation (15). The two sets of estimates are in agreement and are given by $\hat{k} = 8.799$, and $\hat{\theta} = 12.893$. The corresponding bias-corrected estimates of k and 2, as given by equations (16) and (17) are $\hat{k}^{*} = 7.51267$ and $\hat{\theta}^{*} = 15.101$. Note that the bias-corrected MLE of the shape parameter, k = 7.51267, which is quite high; consequently, the skewness of this data set is mild and its MLE = 0.73 (from equation (4)). Goodness-of-fit tests performed on this data set suggest that the data cannot reject the hypothesis that the data are normal or that they are lognormal.

2.3 Computation of UCL of the Mean of a Gamma, G(k,2) Distribution

In the statistical literature, even though procedures exist to compute a UCL of the mean of a gamma distribution (Grice and Bain, 1980, Wong, 1993), those procedures have not become popular due to their computational complexity. Those approximate and adjusted procedures depend upon the Chi-square distribution and an estimate of the shape parameter, k. As seen above, computation of a MLE of *k* is quite involved, and this works as a deterrent to the use of a gamma distribution-based UCL of the mean. However, the computation of a gamma UCL currently should not be a problem due to easy availability of personal computers.

Given a random sample, $x_1, x_2, ..., x_n$ of size n from a gamma, G(*k*,2) distribution, it can be shown that $2n\bar{X} / \theta$ follows a Chi-square distribution, χ^2_{2nk}, with 2nk degrees of freedom (df). It is noted that $(2n\bar{X}) / \theta = 2(X_1 + X_2 + ... + X_n) / 2$. Using a simple transformation of variables, it is seen that each of the random variables, $2X_i/2; i=1,2,...,n$ follows a chi-square, χ^2_{2nk}, distribution. Also those chi-square random variables are independently distributed. Since the sum of the independently distributed chi-square random variables also follows a chi-square distribution, it is concluded that $(2n\bar{X}) / \theta$ follows a chi-square, χ^2_{2nk} distribution with 2nk degrees-of-freedom. When the shape parameter, *k*, is known, a uniformly most powerful test of size " of the null hypothesis, H_0: : $\$ C_s$, against the alternative hypothesis, H_1: : $< C_s$, is to reject H_0 if $\bar{x} / C_s < \chi^2_{2nk}(\alpha) / 2nk$. The corresponding (1-") 100% uniformly most accurate UCL for the mean, :, is then given by the probability statement:

$$P(2nk\bar{x} / \chi^2_{2nk}(\alpha) \geq \mu) = 1 - \alpha, \qquad \textbf{(18)}$$

where $\chi^2_{\upsilon}(")$ denotes the " cumulative percentage point of the Chi-square distribution. That is, if Y follows χ^2_{υ}, then $P(Y \leq \chi^2_{\upsilon}(\alpha)) = \alpha$. In practice, *k* is not known and needs to be estimated from data. A reasonable procedure is to replace k by its bias corrected estimate, $\hat{k}^{*}$, as given by equation (16). This results in the following approximate (1-") 100% UCL of the mean:

$$Approximate\text{-}UCL \cong 2n\hat{k}^{*}\bar{x} / \chi^2_{2n\hat{k}}(a). \qquad \textbf{(19)}$$

It should be pointed out that the UCL given in (19) is an approximate UCL and there is no guarantee that the confidence level of (1-") will be achieved by this UCL. However, it does provide a way of computing a UCL of mean of a gamma distribution. Simulation studies conducted in Section 4 suggest that an approximate gamma UCL thus obtained provides the specified coverage (95%) as the shape parameter, k approaches 0.5. Thus when k$\$$0.5, one can use the approximate UCL given by (19). It should be observed that this approximation is good even for smaller (e.g., n=5) sample sizes.

Grice and Bain (1980) computed an adjusted probability level, $, which can be used in (19) to achieve the specified confidence level of (1-"). For " = 0.05 (confidence coefficient of 0.95), " = 0.1, and " = 0.01, these adjusted probability levels are given below for some values of the sample size n (Table 1). One can use linear interpolation to

obtain an adjusted $ for values of n not covered in the table. The adjusted (1-") 100% UCL of gamma mean, : = k2 is given by:

$$Adjusted\text{ - }UCL = 2n\hat{k}^{*}\bar{x}/\chi^{2}_{2n\hat{k}\cdot}(\beta), \qquad (20)$$

where $ is given in Table 1 for "=0.05, 0.1, and 0.01. Note that as the sample size, n, becomes large, the adjusted probability level, $, approaches ". Except for the computation of the MLE of *k*, equations (19) and (20) provide simple Chi-square-distribution-based UCLs of the mean of a gamma distribution. It should also be noted that the UCLs as given by (19) and (20) only depend upon the estimate of the shape parameter, *k*, and are independent of the scale parameter, 2, and its estimate. Consequently, as expected, it is observed that coverage probabilities for the mean associated with these UCLs do not depend upon the values of the scale parameter, 2. This is further discussed in Section 4.

Table 1. Adjusted Critical Level, $ for Various Values of " and n

	" = 0.05	" = 0.1	" = 0.01
n	probability level, $	probability level, $	probability level, $
5	0.0086	0.0432	0.0000
10	0.0267	0.0724	0.0015
20	0.0380	0.0866	0.0046
40	0.0440	0.0934	0.0070
4	0.0500	0.1000	0.0100

It is observed (Figures 5-7) that except for highly skewed (*k*<0.15) data and samples of small size (e.g., <10), the adjusted gamma UCL given by (20) provides the specified 95% coverage of the population mean. It is also noted that for highly skewed (k<0.15) data sets of small sizes, except for the H-UCL, the coverage probability provided by the adjusted gamma UCL is the highest and is close to the specified level, 0.95. However, for these highly skewed data sets, the H-statistic results in unacceptably large values of the UCL. This is further illustrated in examples 2-4. For values of *k* $0.2, the specified coverage of 0.95 is always approximately achieved by the adjusted gamma UCL given by equation (20), as shown in Figures 7-14.

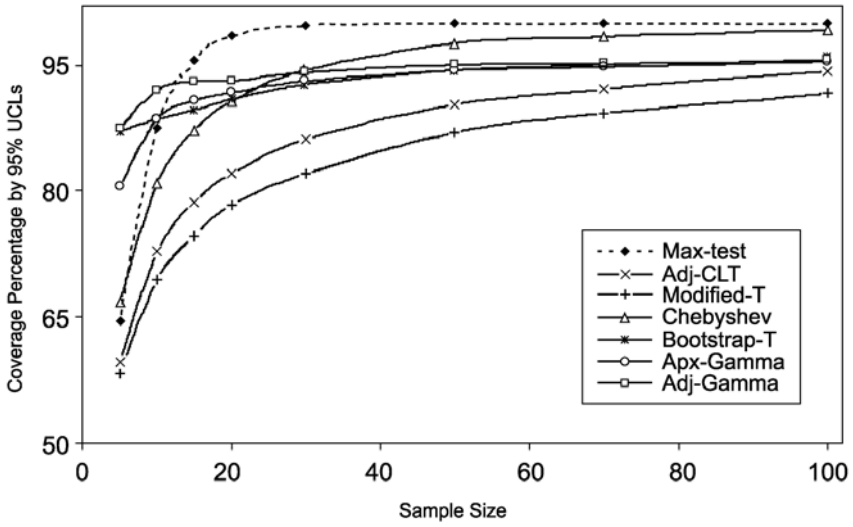

Figure 5. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 0.10, 2 = 50).

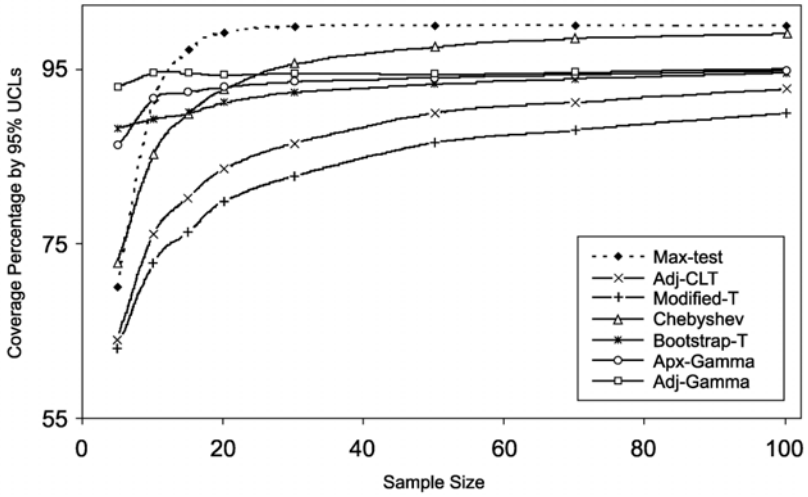

Figure 6. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 0.15, 2 = 50).

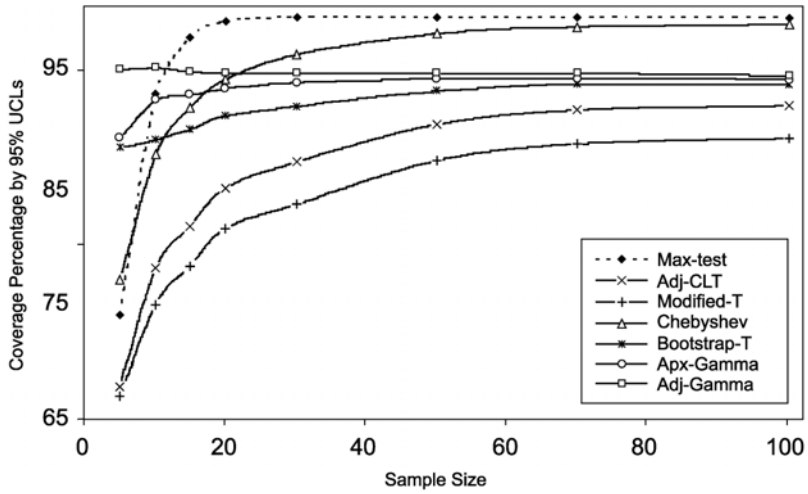

Figure 7. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 0.20, 2 = 50).

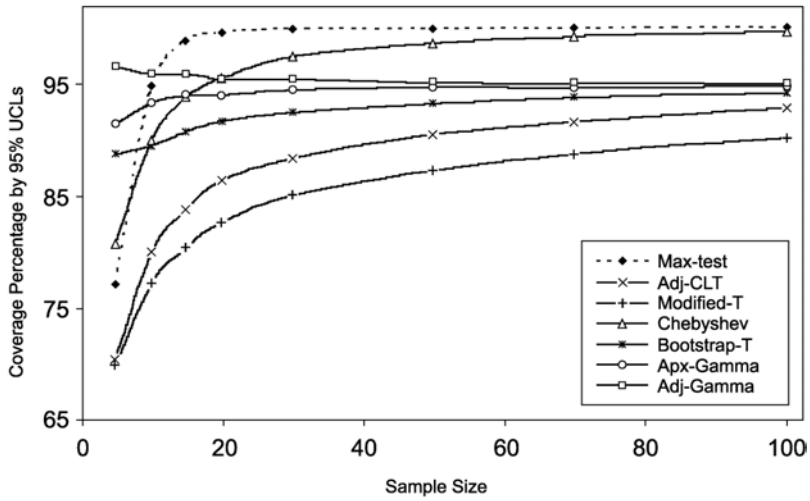

Figure 8. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 0.25, 2 = 50).

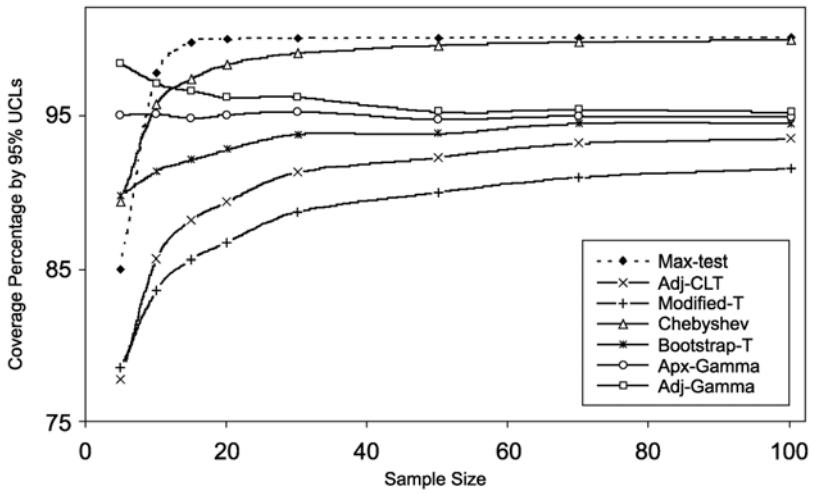

Figure 9. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 0.50, 2 = 50).

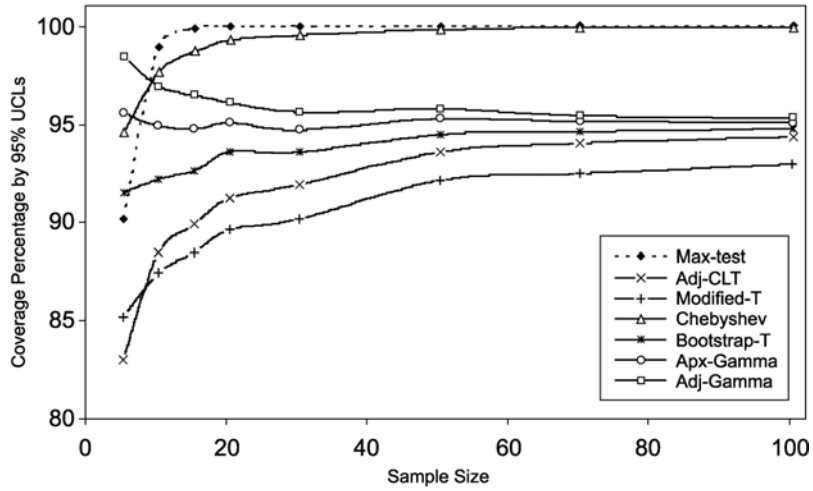

Figure 10. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 1.0, 2 = 50).

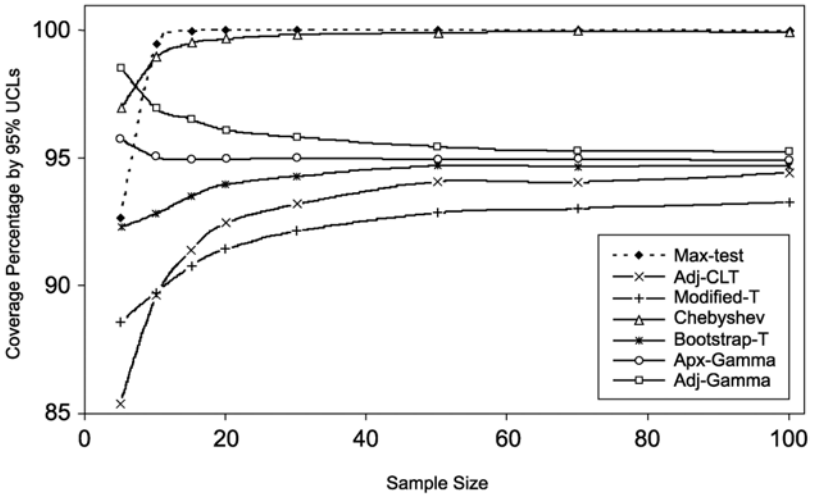

Figure 11. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 2.0, 2 = 50).

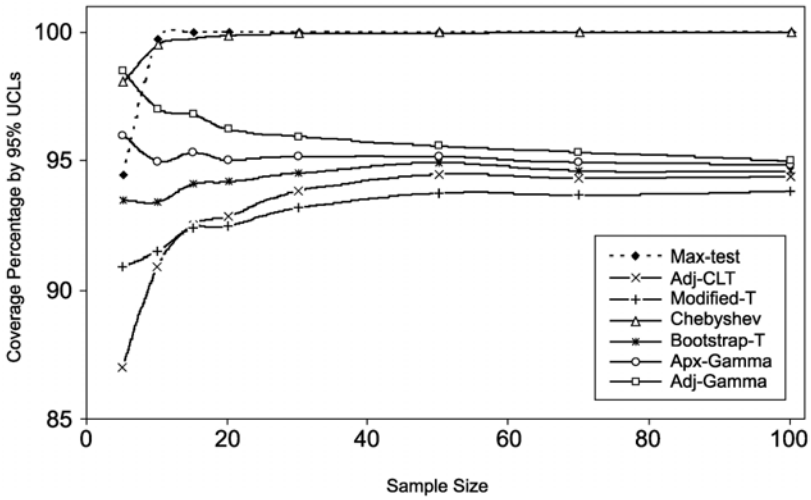

Figure 12. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 4.0, 2 = 50).

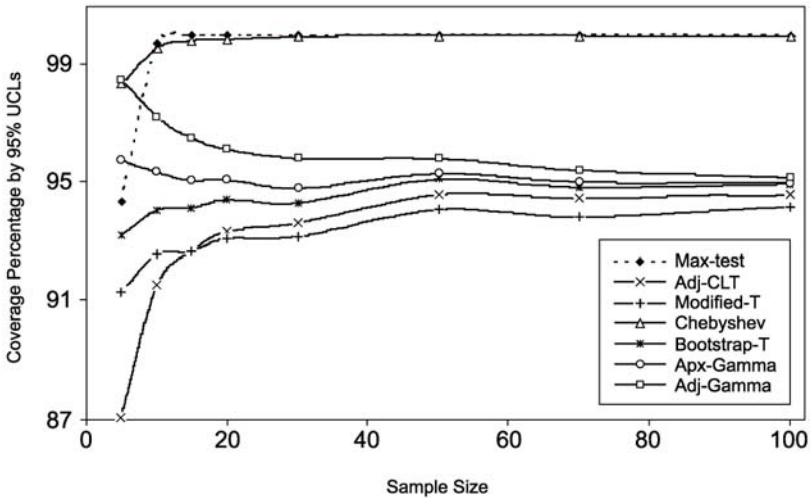

Figure 13. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 6.0, 2 = 50).

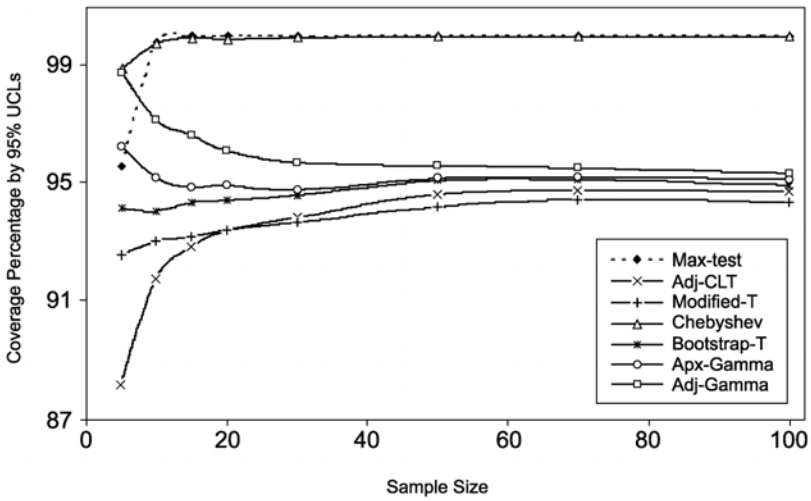

Figure 14. Graphs of coverage probabilities by 95% UCLs of mean of G(k = 10.0, 2 = 50).

Example 1 *(Continued)*
The data set of size 20 and the associated MLEs of parameters k and 2 are given in Example 1. For n=20 and $\propto$ =0.05, the adjusted probability level, $ = 0.038 (Table 1), and the adjusted df, $2n\hat{k}^* = \upsilon^* = 300.507$. The approximate 95% UCL of the mean obtained using equation (19) is given by UCL = 130.447, and the adjusted 95% UCL of mean obtained using equation (20) is given by UCL = 131.901. As noted above, this data set passes both normality as well as lognormality tests. The associated Student's t-statistic based and the H-statistic based UCLs are 127.288 and 134.73, respectively. For this mildly skewed data set, one can use any of these four UCLs.

3. Other UCL Computation Methods

Several authors (Johnson, 1978, Kleijnen, Kloppenburg, and Meeuwsen, 1986, Chen, 1995, Sutton, 1993) have developed inference procedures for estimating the means of asymmetrical distributions. Also, several bootstrap procedures (Efron, 1982, Hall, 1988 and 1992, Manly, 1997) have been recommended for the computation of confidence intervals for means of skewed distributions. These are summarized below and are also included in the simulation experiments described in Section 4. Some examples have been included to illustrate these procedures.

3.1 UCL Based Upon Student's t-Statistic

A (1 ! ") 100% one-sided upper confidence limit for the mean based upon Student's t-statistic is given by the following equation:

$$\text{UCL} = \bar{x} + t_{a,n-1} s_x/\sqrt{n}, \qquad \textbf{(21)}$$

where $t_{",n!1}$ is the upper " th percentile of the Student's t distribution with n !1 degrees of freedom, and the sample variance is given by:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

This UCL should be used when either the data follow a normal distribution, or when the data distribution is only mildly skewed and sample size n is large. For highly skewed data sets, the UCL based upon this method fails to provide the specified (1-") 100% coverage for the population mean, :.

3.2 UCL Based Upon Modified Student's t-Statistic for Asymmetric Distributions

Johnson (1978) and Sutton (1993) proposed the use of a modified t-statistic for testing the mean of a positively skewed distribution. An adjusted (1-") 100% UCL (Singh, Singh, and Engelhardt, 1999) of the mean, :, based upon modified t-statistic is given as follows:

$$UCL = \bar{x} + \hat{\mu}_3/(6ns_x^2) + t_{a,n-1}s_x/\sqrt{n}. \qquad \textbf{(22)}$$

Where, $\hat{\mu}_3$ an unbiased moment estimate (Kleijnen, Kloppenburg, and Meeuwsen, 1986) of the third central moment, μ_3, is given as follows:

$$\hat{\mu}_3 = n\sum(x_i-\bar{x})^3/((n-1)(n-2)). \qquad \textbf{(23)}$$

The simulation study conducted in Section 4 suggests that the UCL based upon the modified-t statistic also fails to provide the specified coverage (95% here) for skewed data sets from gamma distributions.

3.3 UCL of the Mean Based Upon the Adjusted Central Limit Theorem for Skewed Distributions

Given a random sample, $x_1, x_2, \ldots, x_n$ of size n from a population with finite variance, F^2, and 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 : and variance F^2/n. An often cited rule of thumb for a minimum sample size satisfying the CLT is n $ 30. However, this is not adequate if the population is highly skewed (Singh, Singh, and Engelhardt, 1999). A refinement of the CLT approach which makes an adjustment for skewness is discussed by Chen (1995). Specifically, the "adjusted CLT" UCL is given by:

$$UCL = \bar{x} + \left[z_a + \hat{k}_3(1+2z_a^2)/(6\sqrt{n})\right]s_x/\sqrt{n}, \qquad \textbf{(24)}$$

where $\hat{k}_3$, the coefficient of skewness, is given by $\hat{k}_3 = \hat{\mu}_3/s_x^3$. The simulation study conducted in Section 5 suggests that even for larger samples, the adjustment made in the CLT-UCL method is not effective enough to provide the specified

(95%) coverage for skewed data sets. As skewness decreases, the coverage provided by the adjusted CLT-UCL approaches 95% for larger sample sizes, as can be seen in Figures 12-14.

3.4 UCL of the Mean of a Lognormal Distribution Based Upon Land's Method

In practice, a skewed data set can be modeled by both lognormal and gamma distribution. However, due to computational ease, the lognormal distribution is typically used to model such skewed data sets. A (1 ! ")100% UCL for the mean, ∴, of a lognormal distribution based upon Land's H-statistic (1971) is given as follows:

$$\mathrm{UCL} = \exp(\bar{y} + 0.5s_y^2 + s_y H_{1-\alpha}/\sqrt{n-1}) . \qquad (25)$$

where $\bar{y}$ and s_y^2 are the sample mean and variance of the log-transformed data. Tables of values denoted by $H_{1\ !\ "}$ can be found in Gilbert (1987). From the simulation experiments discussed in Section 4, it is observed that H-statistic based UCL grossly overestimates the 95% UCL and consequently, coverage provided by a H-UCL is always larger than the specified coverage of 95%. In Section 4, examples to illustrate this unreasonable behavior of the H-statistic based UCL are included. The practical merit of a H-UCL is doubtful as it results in unacceptably high UCL values. This is especially true for samples of small size (e.g., <25) with values of s_y exceeding 1.5-2.0. This is illustrated in examples 2-4.

3.5 UCL of the Mean Based Upon the Chebyshev Inequality

Chebyshev inequality can be used to obtain a reasonably conservative but stable estimate of the UCL of the mean. The two-sided Chebyshev Theorem states that given a random variable X with finite mean and standard deviation, ∴ and F, we have:

$$P(-j\sigma \le X - \mu \le j\sigma) \ge 1 - 1/j^2. \qquad (26)$$

Here, j is a positive real number. This result can be applied with the sample mean, $\bar{x}$, to obtain a conservative UCL for the population mean. Specifically, a (1-") 100% UCL of the mean, ∴, is given by:

$$UCL = \bar{x} + \sqrt{((1/\alpha) - 1)}\sigma / \sqrt{n} . \qquad (27)$$

Of course, this would require the user to know the value of F. The obvious modification would be to replace F with the sample standard deviation, s_x, but this is estimated from data, and therefore, the result is no longer guaranteed to be conservative. In general, if ∴ is an unknown mean, $\hat{\mu}$ is an estimate, and $\hat{\sigma}$ ($\hat{\mu}$) is an estimate of the standard error of $\hat{\mu}$, then the quantity $UCL = \hat{\mu} + 4.359\ \hat{\sigma}(\hat{\mu})$ will provide a 95% UCL for ∴, which should tend to be conservative, but this is not assured. In this article we use equation (27) to compute a 95% UCL of mean based upon Chebyshev inequality.

From the Monte-Carlo results discussed in Section 4, it is observed that for highly skewed data sets (with k<0.5), the coverage provided by the Chebyshev UCL is smaller than the specified coverage of 0.95. This is especially true when the sample size is smaller than 20. As expected, for larger samples sizes, the coverage provided by the Chebyshev UCL is at least 95%. This means that for larger samples, the Chebyshev UCL will result in a higher (but stable) UCL of the gamma, G(k, 2) mean.

Bootstrap Procedures

General methods for deriving estimates, such as the method of maximum likelihood, often result in estimates that are biased. Bootstrap procedures as discussed by Efron (1982) are nonparametric statistical techniques which can be used to reduce bias of point estimates and construct approximate confidence intervals for parameters such as the population mean. These procedures require no assumptions regarding the statistical distribution (e.g. normal, lognormal, gamma) 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 provide valuable tools for obtaining reliable estimates of the parameters of interest. Use of these methods has been considered in environmental applications (Singh, Singh, and Engelhardt, 1997, 1999; Schulz and Griffin,

1999). Some of those methods are described as follows.

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., $\theta = \mu$) 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 $\bar{x}$. In the bootstrap procedures, repeated samples of size n are drawn with replacement from the given set of observations. The process is repeated a large number of times (e.g., 1000), and each time an estimate, $\hat{\theta}$, of θ (the mean, here) 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 three of those methods are considered: 1) the standard bootstrap method, and 2) bootstrap - t method (Efron, 1982, Hall, 1988), and 3) Hall's bootstrap method (Hall, 1992, Manly, 1997).

3.6 UCL of the Mean Based Upon the Standard Bootstrap Method

Step 1. Let $(x_{i1}, x_{i2}, \ldots, x_{in})$ represent the i^{th} sample of size n with replacement from the original data set $(x_1, x_2, \ldots, x_n)$. Compute the sample mean $\bar{x}_i$ of the i^{th} sample.

Step 2. Repeat Step 1 independently N times (e.g., 1000-2000), each time calculating a new estimate. Denote these 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:

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i - \bar{x}_B)^2}. \qquad (28)$$

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, $\hat{\theta}$.

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

$$t = \frac{\hat{\theta} - \theta}{\hat{\sigma}_B}. \qquad (29)$$

A $(1 - \alpha)$ 100% standard bootstrap UCL for θ, which assumes that equation (29) is approximately normal, is given as follows:

$$UCL = \hat{\theta} + z_a \hat{\sigma}_B. \qquad (30)$$

It is observed that the standard bootstrap method does not adequately adjust for skewness, and the UCL given by equation (30) fails to provide the specified $(1-\alpha)$ 100% coverage of the population mean of skewed data distributions.

3.7 UCL of the Mean Based Upon the Bootstrap – t Method

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 t-statistic, given by (29), directly from data (Hall, 1988). In practice, for non-normal populations, the required t-quantiles may not be easily obtained, or may be impossible to derive exactly. In this method, as before in Steps 1 and 2 described above, $\bar{x}$ 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 ith 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 lower α^{th} quantile of the pivotal quantity (29) is $t_{\alpha,B} = t_{(\alpha N)}$. For example, if $N = 1000$ bootstrap samples are generated, then the 50th ordered value, $t_{(50)}$, would be the bootstrap estimate of the lower 0.05th quantile of the t-statistic as given by (29). Then a $(1-\alpha)$ 100% UCL of mean based upon bootstrap t-method is given as follows:

$$UCL = \bar{x} - t_{(\alpha N)} s_x / \sqrt{n} \qquad (31)$$

3.8 UCL of the Mean Based Upon Hall's Bootstrap Method

Hall (1992) proposed a bootstrap method which adjusts for bias as well as skewness. In this method $\bar{x}_i$, $s_{x,i}$, and $\hat{k}_{3i}$, that is the sample mean, sample standard deviation, and sample skewness, respectively (as given in Section 3.3 above) are computed from the ith resampling (i=1,2,..., N) of

the original data. Let $\bar{x}$ be the sample mean, s_x be the sample standard deviation, and $\hat{k}_3$ be the sample skewness computed from the original data. The quantities W_i and Q_i given as follows are computed for each of the N bootstrap samples, where:

$$W_i = (\bar{x}_i - \bar{x}) / s_{xi}, \text{ and } Q_i(W_i) = W_i + \hat{k}_{3i} W_i^2 / 3 + \hat{k}_{3i}^2 W_i^3 / 27 + \hat{k}_{3i} /(6n).$$

The quantities $Q_i(W_i)$ given above are arranged in ascending order. For a specified (1-") confidence coefficient, compute the ("N)th ordered value, $q_{"}$ of quantities $Q_i(W_i)$. Finally, compute $W(q_{"})$ using the inverse function, which is given as follows:

$$W(q_\alpha) = 3\left(\left(1+\hat{k}_3(q_\alpha - \hat{k}_3/(6n))\right)^{1/3} - 1\right)/\hat{k}_3. \quad \textbf{(32)}$$

Finally, the (1-") 100% UCL of the population mean based upon Hall's bootstrap method (Manly, 1997) is given as follows:

$$UCL = \bar{x} - W(q_\alpha) * s_x. \quad \textbf{(33)}$$

It is observed (Section 4) that the coverage probabilities provided by bootstrap - t and Hall's bootstrap methods are in close agreement. For larger samples these two methods approximately provide the specified 95% coverage to the population mean, k2. For smaller sample sizes, the coverage provided by these methods is only slightly lower than the specified level of 0.95. It is also noted that, for highly skewed (Figures 5-8) data sets (with k#0.25) of small size (e.g., n<10), coverage probability provided by these two methods is higher than the Chebyshev UCL.

4. Examples

Several examples illustrating the computation of the various 95% UCLs of the population mean are included in this section. Software, ProUCL (EPA 2002) has been used to compute some of the UCLs values. Gamma UCLs are computed using the program Chi_test (2002). Examples are generated from the gamma distribution and the lognormal distribution, and UCLs are computed using all of the methods discussed in this paper. It is observed that for small data sets, it is not easy to distinguish between a gamma model and a lognormal distribution. It is further noted that use of a gamma distribution results in practical and reliable UCLs of the population mean. Simulation results discussed in Section 5 suggest that the adjusted gamma UCL approximately provides the specified 95% coverage to the population mean for data sets with shape parameter, k, exceeding 0.1.

4.1 Simulated Examples from Gamma Distribution

Example 2

A data set of size 15 is generated from a gamma, G(0.2, 100), distribution with the true population mean = 20, and skewness =4.472. The data are: 0.7269, 0.00025, 0.0000002548, 0.9510, 0.000457, 32.5884, 0.02950, 1.6843, 3.3981, 170.4109, 59.8188, 0.00042, 0.8227, 0.00726, 2.1037. The data set consists of very small values as well as some large values. These types of data sets often occur in environmental applications. The sample mean is 18.17. Using the Shapiro-Wilk's test, it is concluded that the data also follow a lognormal model. The standard deviation (sd) of log-transformed data is quite large, 5.618; therefore, the H-statistic based UCL of mean becomes unpractically large. The bias-corrected MLEs of k and 2 are 0.16527 and 109.939, respectively. The adjusted (using bias-corrected estimate of k) df, $\hat{\upsilon}^* = 4.958$. For " =0.05, and n=15, the critical probability level, $, to be used is 0.0324 (from Table 1). The UCLs obtained using the various methods are summarized in the following table.

UCL Computation Method	95% UCL of Mean
Approximate gamma UCL (equation (19))	79.968
Adjusted gamma UCL (equation (20))	98.139
UCL based upon t-statistic (equation (21))	38.778
UCL based upon modified t-statistic (equation (22))	40.356
UCL based upon adjusted CLT (equation (24))	47.537
UCL based upon H-statistic (equation (25))	5.4E+13
UCL based upon Chebyshev (equation (27))	69.171
UCL based upon standard bootstrap (equation (30))	36.889
UCL based upon bootstrap - t (equation (31))	102.392
Hall's bootstrap UCL (equation (33))	114.252

Note that the H-UCL becomes unacceptably large. Since the H-UCL exceeds the maximum observed value of 170.41, using the recommendation made in the EPA (1992) RAGS document, one would use that maximum value as an estimate of the EPC term. Simulation results summarized in the next section (Figures 6-7) suggest that for n=15 and an estimate of k = 0.165, the adjusted UCL based upon a gamma model provides the specified 95% coverage to the population mean. Therefore, for this data set, the use of the adjusted gamma UCL of 98.139 (equation 20) is an appropriate choice for an estimate of the EPC term. The maximum observed value represents an overestimate of the EPC term.

Example 3

A data set of size 15 is generated from a gamma distribution with: k=0.5; and 2 =100 with mean, : = k2 = 50, and skewness = 2.828. The data are: 343.31, 102.44, 0.33, 1.42, 13.17, 439.59, 130.66, 158.0, 70.65, 25.05, 144.84, 63.65, 62.50, 11.58, 1.097. Using Shapiro-Wilk's test, it is concluded that these data cannot reject the hypothesis that the data also follow a lognormal distribution with sample mean = 104.553. The bias-corrected estimates of k and 2 are 0.46166, and 226.473, respectively. The adjusted df, $\upsilon^* = 2n\hat{k}^*$, for the Chi-square distribution =13.85. As before, for " =0.05, and n=15, the critical probability level, $ = 0.0324. The 95% UCLs of mean obtained using the various methods described above are given below.

UCL Computation Method	95% UCL of Mean
Approximate gamma UCL (equation (19))	223.879
Adjusted gamma UCL (equation (20))	247.257
UCL based upon t-statistic (equation (21))	163.413
UCL based upon modified t-statistic (equation (22))	165.92
UCL based upon adjusted CLT (equation (24))	175.596
UCL based upon H-statistic (equation (25))	5687.383
UCL based upon Chebyshev (equation (27))	250.22
UCL based upon standard bootstrap (equation (30))	158.798
UCL based upon bootstrap - t (equation (31))	223.665
Hall's bootstrap UCL (equation (33))	461.795

Again note that the H-UCL is 5687.38, which is much higher than the UCLs obtained using any of the other methods. Simulation results suggest that, for n=15 and an MLE of k to be 0.46166, both approximate as well as the adjusted UCLs based upon a gamma model provide the specified 95% coverage to the population mean (Figure 9). Also, note that the Chebyshev UCL is very close to the adjusted gamma UCL. Any of these three methods may be used to compute the UCL of the population mean.

Example 4

A random sample of size n=10 is generated from a gamma (1,100) distribution with mean 100 and skewness=2. The data are: 3.0018, 31.0899, 9.0257, 271.3804, 155.8221, 157.8577, 73.3756, 95.0452, 1.4292, 65.7240. Also, at 0.05 level of significance, these data cannot reject the hypothesis that the data follow a lognormal distribution. They also pass the Shapiro-Wilk's test for normality. The sample mean is 86.375. The bias corrected MLEs of k and 2 are 0.55121 and 156.7006, respectively, and the associated df = 11.0242. For n=10, and " =0.05, the critical probability level, $, to be used (to achieve a confidence coefficient of 0.95) is =0.0267. The UCLs obtained using the various methods are given as follows.

UCL Computation Method	95% UCL of Mean
Approximate gamma UCL (equation (19))	207.435
Adjusted gamma UCL (equation (20))	244.531
UCL based upon t-statistic (equation (21))	136.776
UCL based upon modified t-statistic (equation (22))	138.368
UCL based upon adjusted CLT (equation (24))	141.804
UCL based upon H-statistic (equation (25))	3260.882
UCL based upon Chebyshev (equation (27))	206.222
UCL based upon standard bootstrap (equation (30))	130.526
UCL based upon bootstrap - t (equation (31))	164.356
Hall's bootstrap UCL (equation (33))	148.938

Once again, note that the H-UCL is 3260.882, which is much higher than the UCLs obtained using any of the other methods. Simulation results summarized in the next section suggest that for, for n=10 and an estimate of k to be 0.5512 (Figures 9-10), both the approximate and adjusted UCLs based upon the gamma model at least provide the specified 95% coverage to the population mean. 95% Chebyshev UCL also provides the specified coverage to population mean. For this combination of skewness and sample size, any of these three methods may be used to compute a 95% UCL of population mean.

Example 5

A mildly skewed data set of size 10 was generated from a gamma distribution G(4,100) with mean 400 and skewness =1. The data are 734.9055, 352.2732, 402.2431, 410.0733, 507.1526, 1010.3391, 199.9971, 296.4427, 1241.1702, 392.7091. The sample mean = 554.730. Based upon the Shapiro-Wilk's test, at 0.05 level of significance, the data do not reject the hypotheses of normality as well as of lognormality. The sd of the log-transformed data is 0.561. The bias corrected MLEs of k and 2 are 2.55276 and 217.307, respectively. The associated df =51.055. For n=10, and " =0.05, the critical probability level, $ (to achieve a confidence coefficient of 0.095), to be used is =0.0267. The UCLs obtained using the various methods are given below.

For this data set, the difference between the H-UCL and other UCLs is small. Simulation results suggest that as the sample size increases, these differences in the UCLs will decrease. From these results (Figures 11-12), it is noted that for a sample of size 10 and an estimate of k=2.55, both the approximate Gamma UCL and adjusted gamma UCL at least provide the specified 95% coverage to the population mean. Any of the two methods can be used to compute a 95% UCL of the mean. The Chebyshev inequality results in an overestimate of the UCL.

UCL Computation Method	95% UCL of Mean
Approximate gamma UCL (equation (19))	794.531
Adjusted gamma UCL (equation (20))	847.442
UCL based upon t-statistic (equation (21))	749.638
UCL based upon modified t-statistic (equation (22))	756.657
UCL based upon adjusted CLT (equation (24))	775.617
UCL based upon H-statistic (equation (25))	862.649
UCL based upon Chebyshev (equation (27))	1018.95
UCL based upon standard bootstrap (equation (30))	715.892
UCL based upon bootstrap - t (equation (31))	902.716
Hall's bootstrap UCL (equation (33))	889.773

4.2 Simulated Examples from Lognormal Distributions

Next we consider a couple of small data sets generated from lognormal distributions. It is observed that those data sets also follow gamma models.

Example 6

A sample of size n = 15 is generated from the lognormal distribution with parameters : = 5, F = 2; the true mean of this distribution is 1096.6, the coefficient of variation (CV) is 7.32, and skewness is 414.4. The generated data are: 47.42, 2761.51, 2904.26, 6928.33, 14.73, 7.67, 73.36, 2843.79, 151.71, 103.52, 14.8, 37.32, 24.74, 658.04, 110.42. A goodness-of-fit test showed the data distribution to be non-normal ($P < 0.01$) and also that the data passes the test of lognormality ($P > 0.15$). The software packages ExpertFit (2001) and GamGood (2002) were used to test the goodness-of-fit of the gamma distribution. The observed value of the Anderson-Darling test statistic is 1.094, and the approximate critical value for test size 0.05 is 2.492, and hence an approximate gamma distribution can also be used to model the probability distribution of this data set. The Chi-square goodness-of-fit test with four equal intervals led to the same conclusion. The bias-adjusted estimates of shape, k, and scale, 2, of the gamma distribution are 0.321 and 3466.301, respectively. The 95% UCLs computed from the various methods are given below.

Notice that the H-UCL is more than 5 times higher than the maximum concentration in the sample, and more than 10 times higher than all the other UCLs. All UCLs are larger than the true population mean (1096.6) for this data set. From Figures 8 and 9, it is observed that for an estimate of k=0.321 and n=15, the adjusted gamma UCL = 3276.40 provides the specified 95% coverage to population mean.

Student's t	2005.973
Adjusted CLT	2251.311
Modified t	2053.46
CLT	1946.872
Standard Bootstrap	1917.433
Bootstrap t	2541.425
Hall's Bootstrap	2305.170
Chebyshev (Mean, Std)	3324.25
95% H-UCL	37726.46
Adjusted Gamma UCL	3276.40

Continuing with this example, suppose that another sample is collected and it turns out to be below the detection limit (DL) of the instrument. Suppose further that DL = 10, and following EPA guidance documents, this value is replaced by DL/2 = 5. One would expect that this additional non-detect observation would result in a reduction of the UCL. The UCLs calculated from this sample of n = 16 observations are given below:

Student's t	1884
Adjusted CLT	2122.80
Modified t	1929.28
CLT	1832.02
Standard Bootstrap	1795.20
Bootstrap t	2369.23
Chebyshev	3134.05
H-UCL	40313.2
Gamma UCL	3013.70

The UCLs computed from all but the H-statistic based formula decreased with the addition of one below-detect observation; the H-statistic based formula, however, resulted in a much higher UCL. This is unarguably an unacceptable value.

Example 7

Finally, we consider a data set of size 20 from a highly skewed lognormal model with parameters $:=5$ and $F=3$. For this high value of F, the population mean assuming a lognormal model becomes quite high= 13359.73. In practice, use of such a model will unjustifiably inflate the population mean; therefore, its use to estimate the EPC term is not desirable. Note that the population median is only 148.413. The generated data are: 4453.2441, 337.7879, 2972.0916, 10.4690, 827.7806, 63.2507, 13969.2646, 11.1967, 5.2651, 65.7771, 921.7736, 7.6539, 756.6956, 223.3185, 140.8639, 466.1513, 3.1751, 418.6896, 1.1281, 22.4442. The observations range from 1.1281 to 13969.2646 with sample mean = 1283.9. Note that the population mean is orders of magnitude higher than the sample mean. It is also observed that at 0.05 level of significance, this data set cannot reject the hypothesis of a gamma model. The bias corrected MLEs of shape and scale for a gamma model are 0.28 and 4564.46, respectively. The Kolmogorov-Smirnov (K-S) test statistic for gamma distribution is D= 0.176 which is less than the 5% critical value of about 0.21 (with an estimated shape parameter of 0.28, Table 7, Schneider, 1978) leading to the conclusion that the data cannot reject the hypothesis of a gamma distribution of the data set. The estimated population mean assuming a gamma model is $\hat{k}\hat{\theta}$ = 0.28*4564.46 = 1278.049 which is close to the sample arithmetic mean of 1283.9.

The adjusted 95% Gamma UCL = 3278.41; the 95% UCLs based upon Student's- t and modified- t are 2517.889 and 2616.198, respectively. The 95% Bootstrap-t UCL=5823.31, 95% Chebyshev UCL=4394.61, and 95% H-UCL=87052. As mentioned earlier, use of a lognormal model unjustifiably accommodates large and unpractical values of the mean concentration and its UCLs. From Figures 8 and 9, it is noted that the adjusted gamma UCL will provide the specified 95% coverage to the population mean. Thus, for this data set, a gamma UCL = 3278.41 provides a reasonable estimate of the EPC term.

5. Comparison of the Various UCL Computation Methods

Using Monte Carlo simulation experiments for data sets generated from gamma distributions, the performances of the various UCL computation methods have been compared in terms of the coverage probabilities achieved by the respective UCLs. Similar comparisons (Singh, Singh, Engelhardt, and Nocerino, 2001) have been performed for the various UCL computation methods using data sets generated from lognormal distributions. The methods considered in the present simulation experiments include: Student's t-statistic, modified Student's t-statistic, adjusted CLT, Chebyshev method, H-UCL, approximate-gamma UCL, adjusted gamma UCL, and the three bootstrap methods: standard bootstrap, bootstrap-t method (Efron, 1982; Hall, 1988), and Hall's (1992) bootstrap method. For each of the three bootstrap methods, 1000 resamples have been used. The EPA (1992) RAGS document recommends the use of the maximum observed concentration as an estimate of the EPC term when the H-UCL exceeds the maximum observed value. Therefore, the maximum observed value (called Max-test in this paper) has also been included in the simulation experiments. Thus, 11 EPC computation methods have been considered in these simulation experiments.

The simulation experiments are carried out for various values of the sample size, n = 5, 10, 15, 20, 30, 50, 70, and 100. Random deviates of sample size n were generated from a gamma, G(k,2) population. Various values of k and 2 have been considered. The considered values of k are 0.1, 0.15, 0.2, 0.25, 0.5, 1.0, 2.0, 4.0, 6.0, and 10.0. These values of k cover a wide range of values of skewness, $2/\sqrt{k}$. The simulation experiments were conducted for three values, 1.0, 50.0, and 100.0 of the scale parameter, 2. As noted earlier, gamma distribution based UCLs as given by (19) and (20) only depend upon the estimate of the shape parameter, k and are independent of the scale parameter, 2 and its estimate. Consequently as expected, it is observed that coverage probabilities for the mean

associated with the gamma UCLs do not depend upon the values of the scale parameter, 2, and the differences in the coverage probabilities for these three values of 2 are negligible. Therefore, in this article, the coverage probabilities are graphed for 2=50.0 only. A typical simulation experiment can be described in the following four steps:

Step 1. Generate a random sample of the specified size, n, from a gamma, G(*k*,50) distribution. The algorithm as outlined in Whittaker (1974) has been used to generate the gamma deviates.

Step 2. For each generated sample, compute a 95% UCL of the mean using the various methods described in Sections 2.0, 3.0, and in ProUCL software package (EPA 2002).

Step 3. Repeat steps 1 and 2, 15,000 times.

Step 4. For each UCL computation method, count the number of times the population mean, k2, falls below a respective UCL. The percentages of these numbers provide the coverage probabilities achieved by the various UCL computation methods.

Simulation results suggest that the UCLs based upon Student's-t and standard bootstrap methods fail to provide the specified 95% coverage of the population mean of the distributions considered here. Also, as noted earlier, the H-statistic overestimates the 95% UCL as it provides almost 100% coverage of the population mean. Use of the H-statistic yields impractically large UCL values. This is especially true for highly skewed data sets (k#1). It is also noted that the coverage probabilities provided by bootstrap- t method and Hall's bootstrap method are quite similar. Therefore, the coverage percentages for these four methods: Student's -t, standard bootstrap, Hall's bootstrap, and H-UCL are not included in the graphs presented in this section. The coverage percentages as obtained in Step 4 for the remaining seven (7) methods are given in Figures 5-14. Figures 5-14 have the coverage percentages when k=0.1, 0.15, 0.2, 0.25, 0.5, 1.0, 2.0, 4.0, 6.0, and 10.0, respectively. The following observations have been made from these graphs.

1. From Figures 5-14, it is observed that UCLs based upon the adjusted CLT and modified-t methods fail to provide the specified 95% coverage of the population mean. For mildly skewed data sets (e.g., k$6), the coverage provided by the adjusted CLT-UCL approaches 95% as the sample size becomes larger than 50.

2. It is observed that for highly skewed data with an estimate of k<0.25, even a Max-test (maximum observation) fails to provide the specified 95% coverage of the population mean. This is especially true when the sample size is less than 10 (Figures 5-7). For smaller samples (e.g., #5), the Max-test fails to provide the specified 95% coverage of the mean for values of k as large as 6. Thus, for samples of size 5 or less, the default option of using the maximum observation as an estimate of the EPC term may not be appropriate. It is more appropriate to use an adjusted gamma-distribution-based UCL of the mean. It is also observed that for samples of size 15 or larger, the Max-test always provides at least 95% coverage to population mean. This means for samples of size $15, the Max-test would result in an overestimate of the 95% UCL of the mean.

3. From Figures 5-9, it is also observed that for skewed data sets with k#0.5, the 95% Chebyshev UCL fails to provide the specified 95% coverage of the population mean. This is especially true when the sample size is less 20. When k>0.5, the 95% Chebyshev UCL provides the specified 95% coverage to the population mean even for small samples (Figures 10-14). Furthermore, it is noted that as the sample size increases, the 95% Chebyshev UCL provides at least 95% coverage to population mean resulting in a conservative but stable estimate of the 95% UCL of the mean.

4. It is observed that for highly skewed data sets (Figures 5-6) with k<0.2, and samples of small size (<10), the adjusted gamma UCL provides the maximum coverage (and close to 0.95) of the population mean, and this coverage approaches the specified coverage of 0.95 as k approaches 0.2 (Figure 7). For k$0.2, the adjusted gamma UCL provides at

least 95% coverage of the population mean (Figures 8-14) for samples of all sizes.

5. From Figures 5-8, it is observed that for values of k<0.5, and samples of small size (n<30), the approximate gamma UCL fails to provide the specified 95% coverage of population mean. Also, from Figures 9-14, it is observed that for values of k$0.5, an approximate gamma UCL provides the specified 95% coverage of population mean for samples of all sizes.

6. From Figures 5-14, it is observed that the 95% UCL based upon the bootstrap-t method consistently provides about 90% coverage of population mean. This coverage approaches 95% as the sample size increases.

6. Recommendations

Skewed data sets can be modeled by more than one distribution. Due to the computational ease of working with a lognormal model, users often choose the lognormal distribution for such data sets. It is observed that for small data sets, it is not easy to distinguish between a gamma model and a lognormal distribution. However, there are some fundamental problems associated with the use of a lognormal distribution. The use of a lognormal model unjustifiably elevates the mean and the associated UCL, therefore, its use in environmental applications should be avoided. Since the H-UCL becomes unrealistically large, the Max-test is sometimes used as an estimate of the EPC term. It is shown that for highly skewed (k<0.25) data sets of small size (n<10), the Max-test does not provide the specified 95% coverage of the means of gamma populations and for larger samples, Max-test results in overestimates of the EPC term. Furthermore, the EPC term represents an average concentration in an area, therefore, it should be estimated by a UCL of the mean. In this paper, we have introduced the gamma distribution which is well suited to model highly skewed data sets originating from various environmental applications. It is further noted that use of the gamma distribution results in practical and reliable UCLs of the population mean. Simulation results discussed in Section 5 suggest that the adjusted gamma UCL approximately provides the specified 95% coverage of the population mean for data sets with shape parameter, k, exceeding 0.1. It is, therefore, recommended that for a given data set, the user should use a goodness-of-fit test to see if the data follow a gamma distribution. If the data do follow a gamma distribution, then the user should compute a UCL of the mean based upon a gamma model. It is shown that both approximate and adjusted gamma UCLs behave in a stable manner. For estimated values of the shape parameter, k$0.5, one can use the approximate gamma UCL as an estimate of the EPC term, and for values of k<0.5, one can use the adjusted gamma UCL. Graphs presented in Figures 5-14 cover a wide range of the skewness of gamma distributions. These graphs can be used to determine which method should be used for a given combination of skewness and sample size. For data sets which cannot be modeled by an approximate gamma distribution, one can use a UCL based upon the Chebyshev inequality or the bootstrap t-procedure. These two procedures generally result in conservative, but reasonable, estimates of the EPC term.

References

1. Bowman, K. O., and Shenton, L.R. (1988), Properties of Estimators for the Gamma Distribution, Volume 89. Marcel Dekker, Inc. New York.
2. Chen, L. (1995), Testing the Mean of Skewed Distributions. Journal of American Statistical Association, 90, 767-772.
3. Chi_test Software (2002), A Software that Computes MLEs of Gamma Parameters and an Upper Confidence Limit of the Mean of a Gamma Distribution. Lockheed Martin Environmental Sciences, Las Vegas, Nevada.
4. Choi, S. C., and Wette, R. (1969), Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias. Technometrics, Vol II, pp 383-690.
5. D'Agostino, R.B. and Stephens, M.A. (1986). Goodness-of-Fit-Techniques. Marcel Dekker, Inc., NY.
6. Efron, B. (1982), The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia: SIAM Monograph #38. (Philadelphia: Society for Industrial and Applied Mathematics).
7. EPA (1989), Methods for Evaluating the Attainment of Cleanup Standards, Vol 1, Soils and Solid Media, EPA 230/2-89/042.
8. EPA (1992), Supplemental Guidance to Rags: Calculating the Concentration Term, EPA 9285.7-081, May 1992.
9. EPA (1996), Soil Screening Guidance: Technical Background Document, EPA/540/R-95/128, 9355.4-17A, May 1996.
10. EPA (2002), ProUCL Version 2.1, A Statistical Software, National Exposure Research Lab, EPA, Las Vegas, Nevada, April 2002.
11. ExpertFit Software (2001), Averill M. Law & Associates Inc, Tucson, Arizona.
12. Faires, J. D., and Burden, R. L. (1993). Numerical Methods. PWS-Kent Publishing Company, Boston, USA.
13. GamGood Software (2002), A Goodness-of-fit Testing Software to Test for a Gamma Model. Lockheed Martin Environmental Sciences, Las Vegas, Nevada.
14. Gilbert, R.O. (1987), Statistical Methods for Environmental Pollution Monitoring, New York, Van Nostrand Reinhold.
15. Grice, J.V., and Bain, L. J. (1980), Inferences Concerning the Mean of the Gamma Distribution. Journal of the American Statistical Association. Vol 75, Number 372, pp 929-933.
16. Hall, P. (1988), Theoretical comparison of bootstrap confidence intervals; Annals of Statistics, 16, pp 927-953.
17. Hall, P. (1992), On the Removal of Skewness by Transformation. Journal of Royal Statistical Society, B 54, pp 221-228.
18. Hardin, J.W., and Gilbert, R.O. (1993), Comparing Statistical Tests for Detecting Soil Contamination Greater Than Background. Pacific Northwest Laboratory, Battelle, Technical Report # DE 94-005498.
19. Johnson, N.J. (1978), Modified t-tests and Confidence Intervals for Asymmetrical Populations. Journal of American Statistical Association, 73, pp 536-544.
20. Johnson, N.L., Kotz, S., and Balakrishnan, N. (1994), Continuous Univariate Distributions, Volume 1. Second Edition. John Wiley, NY.
21. Kleijnen, J.P.C., Kloppenburg, G.L.J., and Meeuwsen, F.L. (1986), Testing Mean of an Asymmetrical Population: Johnson's Modified t-test Revisited. Commun. in Statist.-Simula., 15(3), pp 715-731.
22. Land, C. E. (1971), Confidence Intervals for Linear Functions of the Normal Mean and Variance, Annals of Mathematical Statistics 42: pp 1187-1205.
23. Law, A.M., and Kelton, W.D. (2000), Simulation Modeling and Analysis. Third Edition. McGraw Hill, USA.

24. Manly, B.F.J. (1997), Randomization, Bootstrap, and Monte Carlo Methods in Biology. Second Edition. Chapman Hall, London.

25. Schneider, B.E. and Clickner, R.P. (1976). On the distribution of the Kolmogorov-Smirnov Statistic for the Gamma Distribution with Unknown Parameters. Mimeo Series Number 36, Department of Statistics, School of Business Administration, Temple University, Philadelphia, PA.

26. Schneider, Bruce E. (1978), Kolmogorov-Smirnov Test Statistics for the Gamma Distribution with Unknown Parameters. Dissertation, Department of Statistics, Temple University, Philadelphia, PA.

27. Schulz, T. W., and Griffin, S. (1999), Estimating Risk Assessment Exposure Point Concentrations when Data are Not Normal or Lognormal. Risk Analysis, Vol. 19, No. 4, pp 1999.

28. Singh, A.K., Singh, A., and Engelhardt, M. (1997), The lognormal Distribution in Environmental Applications. Technology Support Center Issue Paper, 182CMB97. EPA/600/R-97/006.

29. Singh, A.K., Singh, A., and Engelhardt, M. (1999), Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications. Technology Support Center Issue Paper, EPA/600/S-99/006.

30. Singh, A., Singh, A. K., Engelhardt, M. E., and Nocerino, J. (2001), On the Computation of the Upper Confidence Limit of the Mean of Contaminant Data Distributions. Under EPA Review.

31. Stephens, M. A. (1970), Use of Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics Without Extensive Tables. Journal of Royal Statistical Society, B 32, pp 115-122.

32. Sutton, C.D. (1993), Computer-Intensive Methods for Tests About the Mean of an Asymmetrical Distribution. Journal of American Statistical Association, 88, No. 423, pp 802-810.

33. Thom, H.C.S. (1968), Direct and Inverse Tables of the Gamma Distribution, Silver Spring, MD; Environmental Data Service.

34. Whittaker, J. (1974), Generating Gamma and Beta Random Variables with Non-integral Shape Parameters. Applied Statistics, 23, No. 2, pp 210-214.

35. Wong, A. (1993), A Note on Inference for the Mean Parameter of the Gamma Distribution. Statistics Probability Letters, Vol 17, pp 61-66.

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 subjected to the Agency's peer and administrative review by the EPA and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.