1-1-2000

The gamma distribution as an alternative to the lognormal distribution in environmental applications

Ross Joseph Iaci
University of Nevada, Las Vegas

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

THE GAMMA DISTRIBUTION AS AN ALTERNATIVE TO THE LOGNORMAL DISTRIBUTION IN ENVIRONMENTAL APPLICATIONS

by

Ross J. Iaci

Bachelors in Mathematical Sciences
University of North Carolina at Chapel Hill
1994

A thesis submitted in partial fulfillment
of the requirements for the

Masters of Science Degree
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2000

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

August, 2000

The Thesis prepared by

Ross J. Iaci

Entitled

The Gamma Distribution as an Alternative to the Lognormal Distribution in Environmental Applicatons

is approved in partial fulfillment of the requirements for the degree of

Master of Science

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

PR/1017-53/1-00

ABSTRACT

The Gamma Distribution as an Alternative to the Lognormal Distribution in Environmental Applications

By

Ross J. Iaci

Dr. Ashok K. Singh, Examination Committee Chair
Professor, Department of Mathematical Sciences
University of Nevada, Las Vegas

In environmental applications dealing with data from contaminated sites the positively skewed lognormal distribution has been the most commonly used model. The upper confidence limit (UCL) of the arithmetic mean of a lognormal population is computed by using the H-statistics. Recent concerns have arisen to the effectiveness of the H-Statistic based UCL for the mean of the lognormal distribution in instances of moderately to highly skewed data sets. In this paper the positively skewed Gamma distribution is considered as an alternative to the lognormal distribution and is shown to produce more reasonable UCL's for the mean.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my thesis advisor Dr. Ashok K. Singh for his help in writing this thesis and his help throughout my studies at UNLV.

I would also like to thank Dr. Malwane Ananda, Dr. Rohan Dalpatadu, and Dr. George Miel, who were all of great assistance throughout my studies at UNLV.

In addition I would like to thank Trevor Wilcox for all his technical support, as well as Kim Young and Sue Speakes for their support.

Last and certainly not least, I would like to thank my parents Ross and Cheryl and brother Jason, for all their support through the years.

CHAPTER 1

GENERAL INTRODUCTION

Overview

The purpose of the Comprehensive Environmental Response, Concentration, and Liability Act (CERCLA) is to protect the human health and the environment from contaminants. Additionally, in response to citizen concern over hazardous waste sites, Congress established the Superfund Program in 1980 to locate, investigate and clean up the worst sites nationwide. Sites falling under this mandate are referred to as Superfund Sites. To meet the (CERCLA) mandate the EPA's Office of Emergency and Remedial Response has developed a human health risk assessment process. As a measure of the potential risk associated with a contaminant, the EPA uses an intake equation that contains a source term or concentration term, SAIC [1]. Accurate statistical analysis and interpretation of contaminant concentration data from a Superfund Site is an important concern for the EPA. The EPA Guidance document [2], suggests using the arithmetic average concentration for contaminants based on a set of samples as the source term at Superfund Sites, the reason being that 1) the average concentration is most representative of the concentration that would be contacted at a site over time and 2) toxicity criteria are based on lifetime average exposures.

The concentration term of the intake equation is often the 95 percent upper confidence limit (UCL) for the arithmetic average of the contaminant concentration distribution. In calculating the (UCL), EPA guidance documents [2] suggest that data sets of 10 to 20 samples provide better estimates than data sets of smaller sizes, while sets of 20 to 30 samples provide fairly consistent estimates. The UCL is calculated based upon a t-distribution, which is used when the original data set is considered to be from a Normal distribution. When the original data set does not pass the test for normality, the EPA [2] has suggested using the log transformation of the data. If the log transformed data follows a normal distribution the data is said to be lognormally distributed. The EPA guidance document [2] claims in most cases it is reasonable to assume that samples from Superfund Sites are lognormally distributed. Data from Superfund Sites frequently appear to follow a skewed probability distribution, and in such instances the lognormal distribution is often the suggested model, (see for example the EPA [2]). This seems to be the popular choice since the log transformed data can then be analyzed using normal theory, Linhart [3]. While there has been extensive analysis of the normal distribution, the lognormal distribution has not received the same level of scrutiny.

The H-statistic based upper confidence limit (UCL) for the mean of a lognormal population has been used extensively to make remediation decisions at Superfund sites upon the recommendation of EPA guidance documents, Singh et al [4]. However, doubts on the performance of the H-statistic based UCL have arisen in recent work in environmental statistics. In addition, there have been recent concerns about the effectiveness of the lognormal distribution itself in modeling skewed data sets. Gilbert [5] indicated that statistical tests of hypothesis based on H-Statistics could yield unusually

high instances of wrongful acceptance of the null hypothesis. Singh et al [4], found in cases of skewed and mixed populations the H-Statistic based UCL for the mean was often orders of magnitude larger than the maximum observed data point.

Several other probability distributions are used to model skewed data sets: the Gamma, Chi-Square, Weibull and Exponential distributions, to name a few. However, there is limited use or mention of these distributions in EPA guidance documents. Linhart [3] writes that in dealing with non-negative variables or when the Normal distribution does not appear to fit the data adequately the next obvious choice is the lognormal or Gamma distribution. The Gamma distribution is an appropriate alternative for many reasons, two of which are: 1.) it approaches a normal distribution as its shape parameter becomes large and 2.) the Chi-Square and Exponential distributions are special cases of the Gamma distribution. Moreover, when dealing with small data sets moment estimation based on sufficient statistics utilize the data in the most efficient way, Grice and Bain [6]. Another appealing attribute of the Gamma distribution is that it can be used to model highly skewed to moderately skewed data sets effectively. The limited application of this model in environmental applications is probably due to the mathematical complexities involved in estimating its parameters. With the appropriate numerical procedures and computing power available today these parameters can be easily estimated with a high degree of accuracy.

Another option in the analysis of contaminant concentration data is to employ computer intensive techniques such as the bootstrap method. Bootstrapping procedures involve resampling from the original data set, referred to as nonparametric bootstrapping since no distributional model is assumed, or generating sample sets based upon a

distributional assumption determined from the original data set, referred to as parametric bootstrapping. These methods allow one to tackle a wide range of problems without having to simplify complex problems, Hinkley and Davison [7]. This approach can also be used in more simple problems to check the adequacy of measures obtained and give approximate solutions, Hinkley and Davison [7]. Singh et al [4] showed that even when the data was obtained from a lognormal distribution, upper confidence limits based on the Central Limit Theorem and non-parametric procedures, such as the jackknife and bootstrap out performed the H-Statistic based UCL.

The suitability of a Gamma distribution in modeling site contaminated data as an alternative to the lognormal model is investigated in Chapter 2. In Section 1 of this chapter the Maximum Likelihood Estimates (MLE) for the parameters of the Gamma Distribution are given. The MLE estimates are used since it was shown by Fisher [8] that the efficiency of moment estimates can be as low as 0.22. In the next section the important task of constructing the UCL for the mean of the Gamma distribution is studied. This UCL is obtained from a power study of a uniformly most powerful test conducted by Grice and Bain [6]. In section 3 confidence intervals for the shape parameter of a Gamma distribution are attained from Linhart [3] who modified the confidence limits for the coefficient of variation of the Pearson Type III model. Having these confidence limits for the shape parameter is necessary in the next section, which provides a Goodness Of Fit Test for the Gamma distribution. The Goodness Of Fit Test is used to validate our assumption of an underlying Gamma distribution for contaminant concentration data. The test uses a Kolmogrov-Smirnov (KS) type statistic, which is important since a statistic of this type is not affected by small sample sizes, as is the Chi-

Square Test. The necessary tables and test criteria are given in Schneider [9]. The parametric bootstrapping techniques used in Chapter 2 require one to produce random deviates from a Gamma distribution. An algorithm for producing these deviates is provided in Section 5 from a paper written by Whittaker [10].

In Section 6 the lognormal distribution and various estimates for the mean and UCL of the mean are discussed. The uniformly minimum variance unbiased estimates (MVUE) for moments of the lognormal are presented from a paper written by Bradu and Mundlak [11]. The H-statistic based UCL is introduced based on the work of Land [12][13], as well as the method of using cubic Lagrangian interpolation on the tables of the H-statistic.

The computer intensive methods called bootstrapping are presented in Chapter 2, Sections 7 and 8, and are used as a benchmark in comparing the results obtained in assuming a Gamma distribution or lognormal distribution. Nonparametric bootstrapping techniques and theories used by Davison and Hinkley [7] are explained in Section 7 and numerous confidence intervals are constructed. Among theses are the basic, studentized, and percentile confidence limits. The Bias Correction Skewness (BCA) confidence limits are also offered as a correction to the percentile limits. Parametric bootstrapping methods are the topic of Section 8. Parametric percentile estimates for the UCL of the mean of site contaminated data provide another means for analyzing the appropriateness in assuming an underlying Gamma distribution. For reasons stated by Davison and Hinkley [7] when assuming an underlying Gamma model for site pollutant data the adjusted percentile (BCA) method with skewness correction is used. As a means to better asses the appropriateness of modeling contaminated data sets with a Gamma distribution opposed

to a lognormal distribution the method of calculating the bootstrap p-value based on a likelihood ratio for the statistical hypothesis assuming a null Gamma distribution verse an alternative lognormal distribution is given.

The Results Section of Chapter 2 applies the methods described above on data sets generated from lognormal distributions and actual Superfund Sites to compare the performance of the Gamma model, lognormal model and bootstrapped procedures. All of these estimates are also compared to the estimates based upon the Chebychev, Central Limit, and Adjusted Central Limit Theorems. Additionally, The FORTRAN code written to produce the estimates for the Gamma distribution is benchmarked against results obtained by Schneider [9]. The results show that UCL's based upon the lognormal distribution, especially when the H-statistic is used, often produce overstated estimates for the upper bound of the mean. This situation could lead to not cleaning up a truly contaminated site with the possibility of threatening the environment, humans or both. However, the Gamma model provided upper bounds consistent with the bootstrapping bounds, Chebychev and Central Limit Theorem based bounds, and consistently provided reasonable estimates even when the data set was small or highly skewed.

More detailed analysis of selected topics presented in Chapter 2 are given in Chapter 3, followed by the FORTRAN source code, Chapter 4, used to produce the results of Chapter 2.

References

[1] DOE(1994), East Poplar Creek-Sewer Line Beltway Remedial Investigation. Report Vol 2, Section 5-8, Science Application International Corporation, Oak Ridge, Tennesee, 37831

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

[3] Linhart, H. Approximate Confidence Limits For The Coefficient Of Variation Of Gamma Distributions. Biometrics, Sept., 733. (1965)

[4] Singh, Ashok K., Singh Anita, Englehardt, Max. " The Lognormal Distribution in Environmental Applications," EPA Technology Support Center Issue No. 600/R-97/006, December 1997

[5] Gilbert, Richard O., "Comparing Statistical Tests for Detecting Soil Contamination Greater that Background," Pacific Northwest Laboratory, Technical Report No. DE 94-005498, 1993

[6] Grice, John V., Bain, Lee J. Inferences Concerning the Mean of the Gamma Distribution. Journal of the American Statistical Association, Dec., v. 75, No. 372, 929. (1980)

[7] Davison, A.C., Hinkley, D.V. Bootstrap Methods and their application. Cambridge University Press, New York, NY, 1997.

[8] Fisher, R.A., On the Mathematical Foundations of Theoretical Statistics. Trans. Royal Society London Ser. A 222, p. 309-368

[9] Schneider, Bruce E. Kolmogrov-Smirnov Test Statistics for the Gamma For the Gamma Distribution With Unknown Parameters. Dissertation, Temple University, 1978

[10] Whittaker, J. Generating Gamma and Beta Random Variables with Non-integral Shape Parameters. Appl. Statist, 23, No. 2, 210. (1974)

[11] Bradu, D., Mundlak, Y., Estimation in Lognormal Linear Models. Journal of the American Statistical Society 65 (198-211), 1970

[12] Land, C.E., Confidence Intervals for Linear Functions of the Normal Mean and Variance. Annals of Mathematical Statistics, 42, 11187-1205, 1971

[13] Land, C.E., 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, 1975

CHAPTER 2

THE GAMMA DISTRIBUTION AS AN ALTERNATIVE TO THE LOGNORMAL DISTRIBUTION IN ENVIRONMENTAL APPLICATIONS

This chapter will be presented at the Fourth International Environmetrics Chemometrics Conference, September 18-20, 2000 as an invited paper. The chapter will also be submitted for publication in the Chemometrics and Intelligent Laboratory Systems journal.

Abstract

In environmental applications dealing with data from contaminated sites the positively skewed lognormal distribution has been the most commonly used model. The upper confidence limit (UCL) of the arithmetic mean of a lognormal population is computed by using the H-statistics. Recent concerns have arisen to the effectiveness of the H-Statistic based UCL for the mean of the lognormal distribution in instances of moderately to highly skewed data sets. In this paper the positively skewed Gamma distribution is considered as an alternative to the lognormal distribution and is shown to produce more reasonable UCL's for the mean. However, in using the Gamma distribution numerical algorithms are needed in parameter estimation to compute the UCL. This paper will show that the Gamma distribution is a much more stable model for site remediation data analysis and will provide the necessary numerical procedures needed to estimate a UCL for the mean. Several data sets from Superfund Sites will be used for examples. Data sets from a lognormal distribution will be generated at low, middle and high degrees of skewness, upper bounds from both distributions will be computed and compared to the resampling estimators.

Keywords: Site remediation, Kolmogrov-Smirnov statistics, Goodness Of Fit Testing, Maximum likelihood estimation, Parametric and Non-Parametric bootstrapping techniques, Coefficient of Variation, DiGamma Function, Gamma distribution, scale and shape parameters, Skewness, Upper Confidence Limits

Introduction

This paper is written to provide a more accurate and stable distribution for the modeling of data obtained from sites contaminated with both organic and inorganic pollutants. Most of the literature available on environmental statistics for computing a UCL of the mean contaminant concentration is based on the assumption that the data follows a normal distribution. Often environmental scientists erroneously consider data sets to follow a normal distribution if the coefficient of variance is less than 1, as suggested by a rule of thumb, EPA [1]. Data from Superfund Sites frequently appear to follow a skewed probability distribution, in such instances the lognormal distribution is often the suggested model, (see for example the EPA [2]). This seems to be the obvious choice since the log transformed data can then be analyzed using normal theory, Linhart [3]. While there has been extensive analysis of the normal distribution, the lognormal distribution has not received the same level of scrutiny. However, often the skewness in the data set is caused by other factors such as outliers, biased sampling, multiple populations, or anomalies.

Guidance documents from the EPA have suggested using H-statistics when computing an upper bound for the mean of a lognormal distribution. Details of the H-Statistic can be found in Gilbert [4] and Land [28]. Recent concerns have arisen to the effectiveness of the lognormal distribution itself, and the use of the H-Statistic based UCL for the mean of skewed data sets. Gilbert [5] indicated that statistical testing of hypothesis based on H-Statistics could yield unusually high instances of wrongful acceptance of the null hypothesis. Singh et al [6], found in cases of skewed and mixed populations the H-Statistic based UCL for the mean was often orders of magnitude larger

than the maximum observed data point. The authors also showed that even when the data was obtained from a lognormal distribution, upper confidence limits based on the Central Limit Theorem and non-parametric procedures, such as the jackknife and bootstrap, out performed the H-Statistic based UCL. The worst results were obtained when the standard deviation of the log transformed data starts exceeding 1, and/or the sample size was less than 30. This increase in the standard deviation, the shape parameter of the lognormal distribution, is equivalent to an increase in skewness, since the measure of skewness is defined as $[(e^{\sigma^2}-1)^{1/2}]^3+3[(e^{\sigma^2}-1)]$. Also suggested by the authors was the use of the ML estimates in place of the H-Statistic, which produced results more consistent with the non-parametric upper bounds, Singh et al [6].

Linhart [3] writes that in dealing with non-negative variables or when the normal distribution does not appear to fit the data adequately the next obvious choice is the lognormal or Gamma distribution. The Gamma distribution is often used to model skewed data, but has had little or no mention in environmental applications. This is probably a result of the mathematical complexities involved in estimating its parameters, but with the computing power available today these parameters can be estimated with a high degree of accuracy. The Gamma distribution and its application have been well analyzed, as most of this information is catalogued in the 130 references found in Johnson and Kotz [7]. When dealing with small data sets moment estimation based on sufficient statistics utilize the data in the most efficient way, Grice and Bain [8]. Another appealing attribute of the Gamma distribution is that it can be used to model highly skewed to moderately skewed data sets effectively. Moreover, it can be used when data

sets seem symmetric since it approaches the normal distribution as it's shape parameter becomes large, Grice and Bain [8].

Methods

1. Numerical Methods for Computing the Maximum Likelihood Estimates for the Parameters of the Gamma Distribution

In order to make inferences on the mean of the Gamma Distribution it is first necessary to get efficient estimates of the parameters. It was shown by Fisher [9], that the efficiency of the moment estimates for α and β can be as low as .22, as a result estimates for these parameters will be based on the method of Maximum Likelihood Estimation. The ML estimates of these parameters are known to be asymptotically efficient and consistent, Schneider [10]. Numerical procedures for the ML estimates will be based on the paper by Choi and Wette [11] who provide an algorithm for computation of the parameters and the associated bias. An adjustment for this bias is obtained from a paper by Kotz and Johnson [7]. Assuming the underlying distribution to be Gamma, let $\{x_1, x_2, ..., x_n\}$ be the observed contaminant concentrations from a site then the log-likelihood function is,

$$\ln L(\underline{x}, \alpha, \beta) = (\alpha - 1)\ln \prod_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i / \beta - n \ln \Gamma(\alpha) - n\alpha \ln \beta .$$

Setting the partial derivatives with respect to both parameters equal to zero yields the MLE estimates,

$$\hat{\beta} = \bar{x} / \hat{\alpha} \text{ and } \ln \alpha - \Psi(\alpha) = \ln \bar{x} - \overline{\sum_{i=1}^{n} \ln x_i}$$

Since the MLE for β is dependant on $\hat{\alpha}$ and the observed average contaminant concentration, one only needs to numerically solve for the MLE of α. The equation for the MLE of α is complicated by the presence of the DiGamma Function, $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$. Choi and Wette [11] suggest applying the Newton-Rhapson iteration scheme,

$$\hat{\alpha}_i = \hat{\alpha}_{i-1} - \frac{\ln \hat{\alpha}_{i-1} - \psi(\hat{\alpha}_{i-1}) - Z}{1/\hat{\alpha}_{i-1} - \psi'(\hat{\alpha}_{i-1})}, \text{ where } Z = \ln \bar{x} - \overline{\sum_{i=1}^{n} \ln x_i}$$

where $\hat{\alpha}_i$ represents the ith estimate of α and $\hat{\alpha}_{i-1}$ is the previous estimate respectively. This iterative scheme is convergent for any initial value $\hat{\alpha}_0$ under the constraints $0 \le \hat{\alpha}_0 \le \infty$ and, $\hat{\alpha}\Psi'(\hat{\alpha}) \neq 1$. The initial value of $1/(2\hat{\alpha})$ is suggested since this is the value the expectation of Z approaches, Choi and Wette [11].

In order to apply this iteration scheme the Digamma function, $\psi(\alpha)$, and the Trigamma function, $\psi'(\alpha)$, need to be numerically evaluated at each ith step, this is accomplished using the power series expansions given in Choi and Wette [11],

$$\Psi(\hat{\alpha}) = -\gamma - \hat{\mu}^{-1} + \hat{\alpha}\sum_{i=1}^{\infty}[i(i+\hat{\alpha})]^{-1} \text{ and } \Psi'(\hat{\alpha}) = \sum_{i=0}^{\infty}(i+\hat{\alpha})^{-2}$$

where γ is Euler's constant, 0.57782157. Both the Digamma and Trigamma functions have been analyzed and tabulated in a paper by Pairman [12]. Suggested in Choi and Wette [11], referencing a paper by Jordan [13], approximations to these expansions was shown to be in agreement up to the eight digit of those values obtained by Pairman [12]. The approximations are as follows,

$$\Psi(\hat{\alpha}) \cong \ln \hat{\alpha} - \{1 + [1 - (1/10 - 1/(21\hat{\alpha}^2))/\hat{\alpha}^2]/(6\hat{\alpha})\}/(2\hat{\alpha})$$

and

$$\Psi'(\hat{\alpha}) \cong \{1+\{1+[1-(1/5-1/(7\hat{\alpha}^2))/\hat{\alpha}^2]/(3\hat{\alpha})\}/2\hat{\alpha})\}/\hat{\alpha}$$

In this paper the above iteration schemes were stopped after the absolute difference between $\hat{\alpha}_i$ and $\hat{\alpha}_{i-1}$ was less than 1×10^{-7}.

In a numerical study if the bias of the ML estimates for the parameters of a Gamma distribution Choi and Wette [11] showed that the bias of the estimates decrease as n increases, as expected due to the property of consistency in the ML estimates. More importantly the study gives strong evidence that a positive bias exists, $E(\hat{\alpha}) \geq \alpha$. An adjustment for the bias was given in a paper by Johnson and Kotz [7] as,

$\hat{\alpha} = \hat{\alpha}(1-3/n)+2/3n$ for $n \geq 4$ and $\hat{\alpha} \geq 1$, Schneider [10]

2. Upper Confidence Limit for the Mean of the Gamma Distribution

The most important aspect pertaining to the analysis of contaminant concentration data is to obtain an acceptable limiting value of concentration of contaminants before site remediation is implemented. Again assuming the data $\{x_1, x_2, ..., x_n\}$, follows a Gamma distribution, this value is the upper bound for the mean $\mu = \alpha\beta$. To begin consider the statistic,

$$Z = 2n\bar{x}/\beta \sim \chi^2_{2n\alpha}$$

A $100(1-\rho)\%$ confidence interval for the mean can be constructed in the usual way from the probability statement,

$$P(\chi^2_{\rho(2n\alpha)} \leq 2n\bar{x}/\beta \leq \chi^2_{(1-\rho)(2n\alpha)}) = 1-\rho$$

yielding the upper bound,

$$P(\alpha\beta \leq 2n\bar{x}\alpha/\chi^2_{\rho(2n\alpha)}) = 1-\rho.$$

This of course is not a confidence interval due to the fact that the confidence limits themselves depend on α, which is assumed to be unknown. If we were to replace α with it's MLE estimate $\hat{\alpha}$, we would only have an approximate $(1-\rho)\%$ CI. Due to the sensitive nature and the possible cost of cleanup that could result in not knowing exactly the amount of error introduced by substituting $\hat{\alpha}$, an upper bound will be constructed based on a paper written by J. Grice and L. Bain [8], who estimated this error and give the necessary corrections.

Grice and Bain [8] analyze this problem by investigating the uniformly most powerful tests of the hypothesis,

$Ho: \alpha\beta \geq \alpha\beta_o$ *vs.* the alternative $Ha: \alpha\beta \leq \alpha\beta_o$ and

$Ho: \alpha\beta \leq \alpha\beta_o$ *vs.* the alternative $Ha: \alpha\beta \geq \alpha\beta_o$

where, for the first hypothesis, you reject Ho in favor of Ha if $\bar{x}/\alpha\beta_o \leq \chi^2_{\rho(2n\alpha)}/2n\alpha$, Mood et al [14]. The power or size of the critical region γ for this test is obtained by rearrangement of the above probability statements and defined as,

$$P_1(\alpha,\theta) = P(\bar{x}/\alpha\beta \leq \chi^2_{\theta,(2n\hat{\alpha})}/2n\hat{\alpha}) = \gamma$$

$$P_2(\alpha,\theta) = P(\bar{x}/\alpha\beta \leq \chi^2_{(1-\theta),(2n\hat{\alpha})}/2n\hat{\alpha}) = \gamma$$

where θ is the corrected percentile of the chi-square distribution, when $\hat{\alpha}$ is used in place of α. The term $P_1(\alpha,\theta)$ is of importance because it yields the desired upper bound with a correction for the use of the ML estimate $\hat{\alpha}$. Grice and Bain [8] used Monte Carlo simulation to determine the respective sizes, $P_i(\alpha,\theta)$, by generating random gamma deviates from a Gamma distribution with mean $\mu = 10, \beta = 10/\alpha$ for numerous combinations of α and n. The results of the study are found in Grice and Bain [8].

The limiting values of $P_i(0,\theta) = \lim_{\alpha \to o} P_i(\alpha,\theta)$ and $P_i(\infty,\theta) = \lim_{\alpha \to \infty} P_i(\alpha,\theta)$ are,

$$P_1(0,\theta) = (1 - \ln\theta / n)^{-n+1}$$

$$P_2(0,\theta) = 1 - P_1(0,1-\theta)$$

and

$$P_1(\infty,\theta) = P_2(\infty,\theta) = \int_0^{\infty} \Phi[(v/n)^{1/2} z_\theta] f(v;2,(n-1)/2) dv$$

where $\Phi(x)$ is the standard normal distribution function and z_θ is the standard normal percentile with mean 100 and variance θ, Grice and Bain [8]. These results can be used for interpolation when experiments fall outside the range of the Monte Carlo study.
It was seen that for values of $\alpha \geq 0.15$ that the size of the tests are basically constant, especially in $P_1(\alpha,\theta)$. Due to this Grice and Bain [8] suggest that for most practical purposes at a pre-selected value θ the following is an adequate approximation to the actual power , $\gamma = P_i(\alpha,\theta) \cong P_i(\infty,\theta)$.

For finding a specific $100(1-\rho)\%$ upper bound Grice and Bain [8] provide a table in which the values of θ are given to reach a desired $\rho \cong p_1(\infty,\theta)$, at the levels $\theta = .005,.01,.025,.05,.075,.1,.25$ and $n = 5,10,20,40,\infty$.
An upper bound for the mean is obtained by using the above power of the hypothesis $Ho: \alpha\beta \geq \alpha\beta_o$ *vs.* the alternative $Ha: \alpha\beta \leq \alpha\beta_o$,

$$P_1(\alpha,\theta) = P(\bar{x} / \alpha\beta \leq \chi^2_{\theta,(2n\hat{\alpha})} / 2n\hat{\alpha}) = \gamma$$

giving an upper confidence limit of $2n\bar{x}\hat{\alpha} / \chi^2_{\theta,(2n\hat{\alpha})}$, except for instances where α is known to be small. When α is known to be small Grice and Bain [8] suggest interpolating on θ using the tables containing the actual power of $P_1(\alpha,\theta)$.

3. Confidence Interval for Gamma Shape Parameter α

Now that an approximately unbiased estimate for α has been constructed a confidence interval needs to be obtained for use in Goodness Of Fit Testing. The confidence intervals α are constructed by modifying the confidence intervals for the coefficient of variation, $\alpha^{-1/2}$, obtained by Linhart [3]. By setting $\alpha = \lambda/2$ the Pearson Type III model is obtained and the maximum likelihood estimate for λ is obtained by solving,

$$\ln(\hat{\lambda}/2) - \Psi(\hat{\lambda}/2) = \ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n$$

Linhart [3] notes that for large $\hat{\lambda}$ the left hand side can be estimated by $1/(\hat{\lambda} - 1/3)$. The confidence intervals are determined using Bartlett's [15] approximation,

$$m = \ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n \approx \{[1 + (1+1/n)/3\lambda]/n\lambda\}\chi^2_{n-1}$$

which yields the statistic,

$$n\lambda m/[1 + (1 + n^{-1})/3\lambda] \approx \chi^2_{n-1}$$

The accuracy of this approximation was first investigated by Bishop and Nair [16] and again by Linhart [3] who compared the first four cumulates .01, .05, .95, and .99 to the exact distribution and found for $\lambda = 2, \alpha = 1$, the error to be less than .005 in the first two cumulates and smaller than .001 in the four cumulates for the values, $\lambda \geq 4$, $\alpha \geq 2$. A 100(1-α)% confidence interval is obtained by rearranging the probability statement,

$$P(\chi^2_{(1-\alpha/2),n-1} \leq n\lambda m/[1 + (1 + n^{-1})/3\lambda] \leq \chi^2_{\alpha/2,n-1}) = \alpha$$

$$= P(n\lambda^2 m/\chi^2_{\alpha/2,n-1} \leq n\lambda + (n+1)/3n \leq n\lambda^2 m/\chi_{1-\alpha/2,n-1}) = \alpha$$

and using the quadratic equation setting $\lambda = 2\alpha$ gives,

$$(C_{lower}\chi^2_{(1-\alpha/2),n-1} / 4nm, C_{upper}\chi^2_{\alpha/2,n-1} / 4nm)$$

where

$$C_{lower} = [1 + (1 + 4(n+1)m / 3\chi^2_{(1-\alpha/2),n-1})^{1/2}]$$

$$C_{upper} = [1 + (1 + 4(n+1)m / 3\chi^2_{\alpha/2,n-1})^{1/2}]$$

4. Goodness Of Fit Test for the Gamma Distribution

Now that we have appropriate estimates for the UCL of the mean and parameters of a Gamma distribution we must validate our assumption of an underlying Gamma distribution for soil contamination data. For this we will look at a Kolmogrov-Smirnov (KS) type test statistic for Goodness Of Fit of a Gamma distribution. A KS statistic is one that involves the maximum vertical distance between the empirical distribution function and the assumed distribution function. An advantage to the KS type statistic over the Chi-Square Test is that confidence bands can be formed for the unknown distribution function. One of the most important reasons for the use of a KS type statistic is that is not affected by small sample sizes as is the Chi-Square Test which assumes a large enough sample to provide a good approximation as the distribution of the test statistic, Conover [17]

In testing for Goodness Of Fit we will base it upon the hypothesis that the data follows a Gamma distribution. The testing problem, formally, is to test

$Ho: F(x) = F^*(x)$ *vs* $Ha: F(x) \neq F^*(x)$ where $F(x)$ is the true unknown CDF and $F^*(x)$ is the assumed Gamma CDF. A Kolmogrov-Smirnov type statistic is defined

to be the maximum distance between the empirical distribution function, $S(x) = i/n$, and the assumed CDF $F^*(x)$, i.e.,

$$T_{stat} = Sup_x \left| F^*(x) - S(x) \right|, \text{ Conover [17]}$$

Quantiles of the distribution of this statistic have been well analyzed in Kolmogrov [18] and tabulated by Smirnov [19]. Lee [20] compared the exact power of the Kolmogrov test using a standard parametric test. The test situation was assuming a normal distribution with equal variances to test whether $F^*(x)$ had a mean μ_o vs. some other normal with mean μ_a. The KS test was compared to the normal test and even under the worst conditions the power of the KS test statistic was not much worse than the normal test. Conover [17]

The Kolmogrov Goodness Of Fit Test described above is good for testing when the assumed distribution function is completely specified, meaning parameters of the null distribution function are not estimated. This is not the case when dealing with contaminant concentration data since the parameters of the Gamma distribution need to be estimated. As a result we will used a modification of the above statistic for families of distributions, the statistic type itself remains the same but we will have to use a table other than the table of the quantiles for the Kolmogrov Test Statistic, Conover [17].

The Gamma distribution function estimated from the data will be denoted $F_o(x, \hat{\alpha}, \hat{\beta})$, while the Empirical Distribution Function, EDF, will be defined as,

$$S_n(x) = \sum_{i=1}^{n} \varphi(x - x_i)/n \text{ where } \phi(z) = 1 \text{ for } z \geq 0, 0 \text{ otherwise}$$

The test statistic of the Kolmogrov Smirnov type is defined as,

$$D_n(\hat{\alpha},\hat{\beta}) = Sup_x\left|S_n(x) - F_o(x,\hat{\alpha},\hat{\beta})\right|$$

where $D_n(\hat{\alpha},\hat{\beta})$ is used to denote the dependence of the statistic on the data size and estimated parameter values. In essence we are testing how well the parameters used to define the Gamma distribution fit the observed data. The hypothesis being tested is whether the fitted Gamma distribution is a good fit to the underlying unknown distribution $F(x)$, formally defined as

$Ho: F(x) = F_o(x,\alpha,\beta)$ *vs.* $Ha: F(x) \neq F_o(x,\alpha,\beta)$ for some x, where Ho is rejected for large values of D_n.

The critical values of D_n are not found in the Kolmogrov tables because each $F_o(x,\alpha,\beta)$ to be tested is dependent on estimated parameters, and change with varying data sizes and parameter values. These critical values are provided in a dissertation by Schneider [10]. To obtain the critical values at a specified level ρ Schneider [10] conducted a Monte Carlo study to find the null distribution of $D_n(\hat{\alpha},\hat{\beta})$, by generating a random sample of size n from a $\Gamma(\alpha,1)$ distribution to produce a single $D_n(\hat{\alpha},\hat{\beta})$. The parameter β was set to unity since the distribution of $D_n(\hat{\alpha},\hat{\beta})$ is independent of scale and location parameters, Schneider [10]. This procedure was then repeated 5000 times and the resulting $D_n(\hat{\alpha},\hat{\beta})$ were ordered producing the percentile points.

In the results Schneider [10] observed a "fairly smooth contour" of $D_n(\hat{\alpha},\hat{\beta})$ over α and n at each significance level, and questioned the precision in the third decimal, Schneider [10]. As a fix to this Schneider [10] used a smoothing function of the form,

$$A(\rho)*\left(\frac{e^{B(\rho)\alpha^{-1/2}}}{n^{1/2}+C(p)}\right)$$

at a selected significance level ρ. It was found that this function worked well for $\alpha \geq 0.5$. The smoothed percentage points are found in Schneider [10] and are used for values of α greater than 0.5. For all other values of α the original Monte Carlo results are used. For values of α not in the table Schneider [10] provides the estimated values of the constants to be used in the above function. This table contains piecewise fits with break points between n=4(1)9 and n=10(5)30 which Schneider [10] found to be the most suitable in defining two distinct areas of each contour. In addition this function is suggested for extrapolation purposes when values of n are outside the range of the study.

The study was conducted with data sizes of $n = 10,15,20,25,30$, shape parameters $\alpha = 0.1,0.2,0.5,1,2,5,10,20,50$ at significance levels $\rho = 0.2,0.15,.0.1,0.05,0.01$. In comparison to a paper by Durbin [21], who tabulated the exact null distribution of $D_n(\hat{\alpha},\hat{\beta})$ for exponential data, the maximum deviation in results was 0.004 at $n = 4$ with an average difference less than 0.002, at the significance levels 0.05,0.10 and 0.20. In samples of size 10 and 30 the critical values at the same significance levels never differed by more than 0.001. Also noted for sample sizes of 50 and 100 use of the smoothing function provided similar results,Schneider [10].

The Criteria to determine the appropriate critical values of $D_n(\hat{\alpha},\hat{\beta})$ in testing for the appropriateness of a Gamma distribution are as follows:

1.) Obtain the ML estimates of $\hat{\alpha},\hat{\beta}$ and calculate $D_n(\hat{\alpha},\hat{\beta})$

2.) Calculate the confidence intervals for α, outlined in section 3

3.) Use the smoothed percentage points in Schneider [10], if n or α are not in this table use the smoothing function using the constants tabulated in Schneider [10], and obtain the three critical values corresponding to $\hat{\alpha}, \alpha_{lower}$ and α_{upper}

4.) Reject Ho if $D_n(\hat{\alpha}, \hat{\beta})$ is greater than all three critical values.

Accept Ho if $D_n(\hat{\alpha}, \hat{\beta})$ is smaller than all three critical values.

5.) Otherwise choose a critical value dependant on how conservative the test is made to be.

5. Generating Random Gamma Deviates

In order to use Monte Carlo Simulation or parametric resampling techniques a procedure to generate random Gamma deviates is needed. This is achieved by using the algorithm given by Whittaker [22], which is based on the well-known fact,

$$X = -\beta \sum_{i=1}^{[\alpha]} \ln U_i \approx \Gamma([\alpha], \beta)$$

where each U_i is a random variable from a Uniform distribution between (0,1), and $[\alpha]$ is the integer part of α. The transformations needed to yield the eventual Gamma random variables are given by Whittaker [22] as follows,

letting U_1, U_2 and U_3 be independent $U(0,1)$ variables

define $S_1 = U_1^{1/p}$ *and* $S_2 = U_2^{1/(1-p)}$, where $S_1 + S_2 \leq 1$

and let $Y = S_1/(S_1 + S_2)$, $X_1 = -Y \ln U_3$

then $X_1 \approx \Gamma(p,1)$, and $Y \approx Beta(p, 1-p)$.

Gamma deviates from a distribution with $\beta \neq 1$ are produced by first defining $p = \alpha - [\alpha]$, and letting $X_1 = -\beta Y \ln U_3$, which from Whittaker [22] follows a Gamma distribution with shape parameter p and scale parameter β. Finally, letting $Z = X_1 + X$ yields a random variable from a $\Gamma(p + [\alpha], \beta) = \Gamma(\alpha, \beta)$ distribution.

Numerically the procedure and quality of these random Gamma deviates depends on the manner in which the random variables from a uniform distribution over the unit interval are obtained. The routine for producing the uniform deviates in this paper is found in Numerical Recipes in Fortran Second Edition, Press et al [23]. This routine was chosen because the shuffle used produces low-order serial correlations.

6. The Lognormal Distribution and the H-Statistic

Due to the EPA's continued and suggested use of the lognormal distribution for making inferences on data collected from polluted sites, the H-Statistic and other estimates will be given in this section. The contaminated data follow a lognormal distribution if the logarithmic transformation of the data follows a normal distribution. Moreover, it is obvious from the definition that the appropriateness in assuming a lognormal distribution is dependent upon the proper measurement of normality of the log transformed data. Mathematically the lognormal distribution in environmental applications is represented as follows,

Letting $X = \{x_1, x_2, ..., x_n\}$ be the data collected from a contaminated site if $Y = \ln X$ follows a $N(\mu, \sigma^2)$ distribution then $X = e^Y$ follows a $LN(\mu, \sigma^2)$ distribution

with the pdf, $f_x = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(\ln x-\mu)^2} \frac{1}{x} dx$. Although it is common practice to symbolize the mean and variance of both distributions with μ and σ^2, in this paper μ_1 and σ_1^2 will denote the mean and variance of the original log transformed data. The moments of the lognormal distribution are found by considering the well known moment generating function of Y,

$$M_y(t) = E(e^{ty}) = e^{\mu t+\sigma^2 t^2/2} \text{ which implies that } E(Y^k) = E(e^{ky}) = e^{\mu k+\sigma^2 k^2/2}$$

Therefore the mean and variance are,

$$\mu_1 = e^{\mu+\sigma^2/2}$$

$$\sigma_1^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

Other parameters of interest are,

$$\text{Coefficient of Variance} = CV = \sigma_1 / \mu_1 = (e^{\sigma^2} - 1)^{1/2}$$

$$\text{Measure of skewness} = (CV)^3 + 3(CV)$$

In using this distribution to model site contaminated data, finding an upper bound for the mean contaminate concentration will again be the main focus. The first estimate to be considered is the sample mean $\bar{x}$, which is an unbiased estimate regardless of the underlying distribution. Moreover, if the underlying distribution is normal then the sample mean is the MVUE estimate. In the case when the underlying distribution is lognormal $\bar{x}$ is not the MVUE estimate. The ML estimates are obtained by using the estimates $\bar{y} = \sum_{i=1}^{n} \ln x_i / n$ and $s_y^2 = \sum (\ln x_i - \bar{y})^2 / n$ in the above equations giving,

$$\hat{\mu}_1 = e^{\bar{y}+s_y^2/2} \text{ and } \hat{\sigma}_1^2 = e^{2\bar{y}+s_y^2}(e^{s_y^2} - 1)^{1/2}$$

It is suggested that these estimates be used when the CV is less than 1.2, i.e. the data is not heavily skewed, Gilbert [4] and Koch and Link [24]. If the CV is greater than 1.2 Gilbert suggests using the MVUE estimated derived by Finney and presented in a paper by Bradu and Mundlak [25]. The MVUE solution derived as follows,

Let z, s^2 be two independent random variables where $z \sim N(\xi, \upsilon\sigma^2)$ and $s^2 \sim \sigma^2 \chi_n^2 / n$. The moments of s^2 are defined as $E(\sigma^2 \chi_n^2 / n)^k$ and can be represented in the functional form as follows,

$$E(s^{2k}) = \frac{n(n+2)...(n+2k)}{n^k (n+2k)} \sigma^{2k} \quad k = 0,1,.....\infty$$

Finney [26] introduced the following function,

$$g_n(t) = \sum_{k=0}^{\infty} w_k(t) = \sum_{k=0}^{\infty} \frac{n^k (n+2k)}{n(n+2)...(n+2k)} \left(\frac{n}{n+1} \right)^k \frac{1}{k!} t^k$$

which is combined with the previous equation and algebraically simplified to,

$$E(g_n(As^2)) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{n}{n+1} A\sigma^2 \right)^k = e^{nA\sigma^2/(n+1)}$$

$\therefore g_n(As^2)$ is an unbiased estimate for $e^{nA\sigma^2/(n+1)}$.

Using the estimates, $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \sim \chi_{n-1}^2$ and $\bar{x} \sim N(\mu, \sigma^2 / n)$ the following is given by Bradu and Mundlak [25],

$$E[e^{\bar{x}} g_n \left(\frac{n}{n-1} (c - 1/2n\tau^2) s^2 \right)] = E[e^{\bar{x}} g_n \left(\frac{2nc - \tau^2}{2(n-1)} s^2 \right)] = e^{\tau\mu + \sigma^2 c}$$

$\therefore$ $e^{\bar{x}} g_n \left(\frac{n}{n-1} (c - 1/2n\tau^2) s^2 \right)$ is an unbiased estimate for $e^{\tau\mu + \sigma^2 c}$

This estimator is also the uniformly minimum variance unbiased estimate (MVUE), since s^2 and $\bar{x}$ are jointly sufficient and complete statistics, Bradu and Mundlak [25].

So for the lognormal distribution we have, assuming $Y = \ln X$ follows a normal distribution,

$$\mu_1 = e^{\mu+1/2\sigma^2} \quad \Rightarrow \quad \hat{\mu}_1 = e^{\bar{y}} g_n(1/2s^2) \ , \ \tau = 1 \text{ and } c = 1/2$$

$$\text{Median} = e^{\mu} \quad \Rightarrow \quad e^{\bar{y}} g_n\left(-\frac{1}{2(n-1)}s^2\right), \ \tau = 1 \text{ and } c = 0$$

$$\sigma_1^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \ \Rightarrow \quad \hat{\sigma}_1 = e^{2\bar{y}}\left[g_n(2s^2) - g_n\left(\frac{n-2}{n-1}s^2\right)\right]$$

In this paper a program evaluates the series $g_n(t)$, tables of these values can be found in any of the following references: Aitchison and Brown [27], Koch and Link [24], or Gilbert [4]. The variance of the estimate, $\hat{\mu}_1$ is also given by Bradu and Mundlak [25] as,

$$\hat{\sigma}^2_{\hat{\mu}_1} = e^{2\bar{y}\left[\left(g_n(s_y^2/2)\right)^2 - g_n\left((n-2)s_y^2/(n-1)\right)\right]}$$

To find an upper bound for the mean of the lognormal distribution, consider the *pth* quantile for the distribution of the random variable X defined as $P(X \le x_p) = p$.

Moreover, the *pth* quantile of a standard normal random variable is,

$$P(Z \le z_p) = P\left(\frac{x-\mu}{\sigma} \le z_p\right) = P(x \le \mu + z_p\sigma) = p$$

therefore the *pth* quantile for a random variable X of a lognormal distribution is given by $x_p = e^{\mu+z_p\sigma}$. So an upper bound for the mean using the above MVUE estimates is

$$\mu_1^{upper} = e^{\hat{\mu}_1 + z_p\hat{\sigma}_1}$$

Another estimate that can be used is to place the ML estimates, $\bar{y}$ and $s_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / n$, into the above equation. A disadvantage to using these estimates is the lognormal mean and percentiles are biased. The estimate $s_y = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}$ can also be used with only a slight difference numerically.

The most commonly used $100(1-\alpha)\%$ UCL for the mean of the lognormal is based on the H-statistic given by Land [28] [29] as

$$\mu_1^{upper} = e^{\left(\bar{y} + 1/2 s_y^2 + \frac{s_y H_{1-\alpha}}{(n-1)^{1/2}}\right)}$$

where $s_y = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}$ and the values of $H_{1-\alpha}$ are tabulated in Land [29] based on s_y, n and α. The UCL based on the H-statistic is the recommended method for skewed data, and has optimal properties if the underlying distribution is truly lognormal, Gilbert [4]. However, in practice this method can produce poor and misleading results if the data contains outliers or is a mixture of distributions Singh et al [6].

When values of s_y and n are not represented in the tables Land [29] suggests the use of Lagrangian cubic interpolation. Abramowitz and Stegun [30] give the functional form as,

$$f(x_o + ph) = A_{-1} f_{-1} + A_0 f_0 + A_1 f_1 + A_2 f_2 + R_3 \approx \frac{-p(p-1)(p-2)}{6} f_{-1} + \frac{(p^2-1)(p-2)}{2} f_0 - \frac{p(p+1)(p-2)}{2} f_1 + \frac{p(p^2-1)}{6} f_2$$

The cubic Lagrangian interpolation is given as follows, Gerald and Wheatley [31],

$$P_3(x)=\frac{(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)}f_1+\frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}f_2+$$
$$\frac{(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)}f_3+\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_1)(x_4-x_2)(x_4-x_3)}f_4$$

In cases where s_y and n are both not given in the table the above formula will be adapted by performing cubic Lagrangian interpolation in the x direction followed by cubic Lagrangian interpolation in the y direction. This involves holding y constant and interpolating on the x values at each of the four values followed by interpolation on the y values holding x constant.

For further comparisons a conservative upper bound for the mean is obtained by applying the Chebychev theorem with the above MVUE estimates, $\hat{\mu}_1=e^{\bar{y}}g_n(1/2s^2)$ and estimated variance $\hat{\sigma}^2_{\hat{\mu}_1}=e^{2\bar{y}\left[\left(g_n(s_y^2/2)\right)^2-g_n\left((n-2)s_y^2/(n-1)\right)\right]}$, yielding a $100(1-\rho)\%$ Chebychev UCL

$$\mu_1^{upper}=\hat{\mu}_1+\rho^{-1/2}\hat{\sigma}_{\hat{\mu}_1}\text{, for a desired significance level } \rho,\ k=\rho^{-1/2}$$

In general this estimate should tend to be conservative but is not assured.

7. Nonparametric Bootstrapping

The results of the UCL based on the Gamma distribution are compared with those based on the lognormal distribution by using the computer-intensive method of bootstrapping. The method of bootstrapping simulates repeated samples by randomly picking a data point from the original data set with equal probability. From these created data sets various methods will be used to obtain a UCL of the mean based on the Empirical Distribution Function (EDF) of each set.

In the following discussion, from Davison and Hinkley [32], T is the statistic on which to base inferences regarding the unknown parameter θ, and $t=T(x_1, x_2,\ldots,x_n)$ is the value of T calculated from the sample $\{x_1, x_2,\ldots,x_n\}$. The first and most basic bootstrap estimate is to assume that T follows a normal distribution with mean $\theta + \beta$ and variance v, where β is the bias. The bias and variance are assumed unknown and can be estimated by the bootstrap method by taking N simulated data sets and computing the *pth* quantile estimate of T - θ. This is done letting T_i^* be the ith simulated data set and t_i^* be the estimate calculated from T_i^*, then calculate and order $t_i^* - t$. The *pth* quantile estimate of T - θ is then defined as $t_{Np}^* - t$, where N is chosen such that Np is an integer. The $100(1-\alpha)\%$ basic bootstrap confidence interval is,

$$(t-(t_{N(1-\alpha/2)}^* - t),\ t-(t_{N(\alpha/2)}^* - t))$$

The accuracy of these estimates depends upon the agreement of the distribution of $T^* - t$ to that of T - θ. Davison and Hinkley [32] suggest that if the distribution of T - θ is dependent on unknowns, then alternative expressions contrasting T and θ should be used. One way to do this is to use studentized comparisons, which is given by estimating, $S = \frac{T-\theta}{v^{1/2}}$ with the bootstrap estimate, $S^* = \frac{(T^* - t)}{V^{*1/2}}$, where T^* and $V^{*1/2}$ are based on the simulated sample $X_1^*, X_2^*,\ldots, X_N^*$. Again the values of $s_i^* = \left(\frac{t_i^* - t}{v_i^{*1/2}}\right) i = 1,2,\ldots,N$ are ordered with the *pth* quantile estimate being s_{Np}^*. Substituting the simulated quantile estimate of S into the well known studentized confidence intervals generates the studentized bootstrap confidence limits, also referred to as bootstrap-t limits,

$$(t - v^{1/2} s^{*}_{N(1-\alpha/2)},\ t - v^{1/2} s^{*}_{N\alpha/2})$$

Generally if the sample is large enough a simple approximation using the quantiles of the standard normal distribution can be used. For contaminant concentration data this is often inadequate and is calculated in this paper for comparative purposes. The improvement gained over the normal approximation will be measured by looking at a $Q-Q$ plot of the t^{*} values.

Often in parametric analysis statistics T can be used to estimate θ for which approximate distributions exist and can be extended to nonparametric analysis. If this is the case approximate distributions for T can be calculated using the delta method, which yields the delta method variance estimate, Davison and Hinkley [32].

Mathematically the influence values are obtained by looking at the influence function,

$$L_t(x;F) = \lim_{\varepsilon \to 0} \frac{t[(1-\varepsilon)F + \varepsilon Hx] - t(F)}{\varepsilon} = \frac{\partial t[(1-\varepsilon)F + \varepsilon Hx]}{\partial \varepsilon} \text{ at } \varepsilon = 0$$

where H is the unit step function going from 0 to 1 at $\mu = x$. Substituting $\hat{F}$ for F results in the empirical influence function $l(x) = L_t(x;\hat{F})$. Assuming a smooth function g the estimate of the unknown scalar parameter θ can be put in the linear form,

$$t(G) \cong t(F) + \int L_t(x;F)\,dG(x)$$

Taking $G = \hat{F}$, and applying the first order approximation gives the nonparametric delta method,

$$t(\hat{F}) \cong t(F) + \int L_t(x;F)\,d\hat{F}(x) = t(F) + 1/n \sum_{j=1}^{n} L_t(x_j;F)$$

Next applying the central limit theorem to the right hand side of the equation implies that $T-\theta$ follows an approximate normal distribution with mean 0, since $\int L_t(x;F)dF(x)=0$, and variance $v_L(F)$. The variance is mathematically written,

$$v_L(F)=1/n\,\mathrm{var}(L_t(X))=1/n\int L_t^2(X)dF(x)$$

and is approximated by substituting $\hat{F}$ for F giving the nonparametric delta method variance estimate,

$$v_L=v_L(\hat{F})=1/n^2\sum_{j=1}^{n}l_j^2$$

where $l(x)=l(x_j)$ are the values of the empirical influence function $l(x)=L_t(x;\hat{F})$, referred to as the empirical influence values.

Davison and Hinkley [32] note that making a bias adjustment in the numerator of $S=\dfrac{T-\theta}{v_L^{1/2}}$ is rarely effective and is implicitly made in the bootstrapped estimate,

$$S^*=\frac{(T^*-t)}{V_L^{*1/2}}.$$

The studentized bootstrap estimate using the nonparametric delta method variance is,

$$(t-v_L^{1/2}s^*_{N(1-\alpha/2)},\ t-v_L^{1/2}s^*_{N\alpha/2})$$

Moreove, the influence values need to be calculated based on the EDF of each of the simulated data sets,

$$v_L^*=1/n^2\sum_{j=1}^{n}l^2(x_j^*;\hat{F}^*)\ \text{where}\ l(x_j^*;\hat{F}^*)\cong l(x_j^*\hat{F})-1/n\sum_{j=1}^{n}l(x_j^*;\hat{F}^*)$$

In analyzing contaminant concentration data nonparametrically, we use $t=\bar{x}$ or $t(F)=\int xdF(x)$. The influence function is derived taking

$t[(1-\varepsilon)F + \varepsilon H_x] = (1-\varepsilon)\mu + \varepsilon x$ and differentiating,

$L_t(x) = \dfrac{\partial[(1-\varepsilon)\mu + \varepsilon x]}{\partial \varepsilon}$, evaluated at $\varepsilon = 0$ equals $x - \mu$.

So the empirical influence function becomes, $l(x) = x - \bar{x}$ and is evaluated by $l_j = x_j - \bar{x}$.

This results in the nonparametric delta method variance estimate,

$v_L = v_L(\hat{F}) = 1/n^2 \sum_{j=1}^{n} (x_j - \bar{x})^2$, Davison and Hinkley [32]

The next nonparametric bootstrap estimate useful in analyzing contaminant concentration data is the basic percentile method and the adjusted percentile method, where the adjustment made is for skewness. A possible improvement to a bootstrap interval may lie in finding a transformation or looking at the simulation results to gain insight as to what might be a suitable transformation. Percentile methods differ in that they implicitly use the existence of a good transformation without having to find the transformation. Davison and Hinkley [32] explain this mathematically as follows:

Define the unknown transformation of T as $U = h(t)$, which has a symmetric distribution. Then find the basic bootstrap interval for the transformed parameter defined as $\phi = h(\theta)$ with significance level α. Assuming the transformation on T follows a symmetric distribution implies the quantiles, $U - \phi$, follow a symmetric distribution. Moreover, the following is true,

$$P(U - \phi \le a_{\alpha/2}) = P(U - \phi \ge a_{1-\alpha/2}) = \alpha/2$$

which implies,

$$P(\phi \ge U - a_{\alpha/2}) = P(\phi \le U - a_{1-\alpha/2}) = \alpha/2 \Rightarrow$$

$$P(U - a_{\alpha/2} \le \phi \le U - a_{1-\alpha/2}) = \alpha \Rightarrow$$

$$P(U - a_{1-\alpha/2} \le \phi \le U - a_{\alpha/2}) = 1 - \alpha$$

therefore the $100(1-\alpha)\%$ confidence interval is $U - a_{1-\alpha/2}, U - a_{\alpha/2}$. The basic bootstrap interval would then be written as $u - u^*_{N(1-\alpha/2)}, u - u^*_{N\alpha/2}$, which is the same as $u^*_{N\alpha/2} - u, u^*_{N(1-\alpha/2)} - u$, due to symmetry. Substituting into the basic bootstrap confidence limits given earlier as $t - (t^*_{N(1-\alpha)} - t)$, $t - (t^*_{N\alpha} - t)$ gives,

$u - [u - u^*_{N\alpha}], u - [u - u^*_{N(1-\alpha/2)}]$, which equals

$u^*_{N\alpha}, u^*_{N(1-\alpha/2)}$

Transforming back gives the bootstrap percentile estimate,

$t^*_{N\alpha}, t^*_{N(1-\alpha/2)}$

which is independent of h, Davison and Hinkley [32].

The ability of the percentile method to provide good estimates depends upon T being unbiased on the transformed scale, which is rarely the case. Another problem is the shape of the distribution of T changes as the sampling moves from F to $\hat{F}$. The adjusted percentile method can be used to overcome the difficulties of the percentile method. The adjusted percentile method is obtained by applying the method used in the parametric case with no nuisance parameters, since no underlying model is assumed. The nonparametric exponential family is constructed using the least favorable family of the multinomial distribution of the frequencies of the resampled data. The resampling model used to obtain a resampled random variable X^*, is the exponential tilted distribution

$$P(X^* = x_i) = p_i = \frac{e^{\eta l_i}}{\sum_{j=1}^{n} e^{\eta l_i}}$$

where l_i is the empirical influence value of t at x_i. The function η is a monotone function of the parameter θ with inverse $\eta(\theta)$. The bias correction factor is $w = \Phi^{-1}[\hat{G}(t)]$, where Φ implies the standard normal distribution, and is estimated with nonparametric bootstrap simulation as,

$$\hat{w} = \Phi^{-1}\left[\frac{\#(t_N^* \le t)}{N}\right]$$

The skewness correction factor is obtained by substituting the nonparametric analogue into the no nuisance parametric equation, yielding,

$$a = 1/6 \frac{\sum l_i^3}{(\sum l_i^2)^{3/2}}$$

The Bias Correction Skewness (BCA) confidence limits are then given by,

$$(t_{N\alpha}^*, t_{N(1-\alpha/2)}^*), \text{ where } \breve{\alpha}/2 = \Phi\left[w + \frac{w + z_{\alpha/2}}{1 - a(w + z_{\alpha/2})}\right]$$

8. Parametric Bootstrapping Assuming a Gamma Distribution

The following discussion on parametric bootstrapping methods is based on the work form Davison and Hinkley [32]. Parametric bootstrap resampling consists of obtaining the parameter estimates discussed above and then generating N data sets of size equal to the original data set from an assumed distributional model. As in the previous section we are interested in the properties of the distribution of T which yields, t, an estimate of the parameter θ. Moreover, we will assume the random sample of the contaminant concentration data $X = \{x_1, x_2, ..., x_n\}$ follow a Gamma distribution with CDF $F(x)$. The parameters of the Gamma distribution will be estimated using the MLE method given

earlier to obtain the fitted Gamma model with CDF, $\hat{F}(x)$. The simulated data sets, $X_i^* = \{x_1^*, x_2^*, ..., x_n^*\}$, $i = 1,2,...,N$, are generated according to the fitted model. Further, the *ith* parameter estimate of θ, calculated from the *ith* simulated data set, is denoted T_i^*. All of the nonparametric bootstrap confidence interval estimates defined in the previous section can be used simply by substituting in the Gamma estimates for the mean and the simulated estimates respectively. More sophisticated tests such as the bootstrap p-value for testing of distributional families and the adjusted percentile method for the UCL are given in this section.

To better asses the appropriateness of modeling contaminant concentration data , the bootstrap p-value is calculated with the null hypothesis H_o of a Gamma distribution family verse the alternative H_a of a lognormal family. The test statistic T_{stat} measures the difference between the information contained in the data and the null hypothesis, where t is the observed value observed from the data. Large values of T_{stat} will result in rejection of the null hypothesis in favor of the alternative. The associated p-value or significance probability measures the probability of observing a value of T_{stat} under the null hypothesis. Therefore large p-values lead to acceptance of H_o. Mathematically this is written, $p = P(T_{stat} \geq t | H_o)$.

Even when the null distribution of T_{stat} does not depend on nuisance parameters, often obtained by standardizing or conditioning on the sufficient statistics, calculating the p-value may be difficult or even impossible, Davison and Hinkley [32]. When these situations arise the Monte Carlo resampling method provides good approximations, Davison and Hinkley [32]. Monte Carlo tests obtain the observed test statistic t

from N simulated data sets under the null hypothesis distribution. These values, denoted by $t_1^*, t_2^*, ..., t_N^*$, are ordered to obtain the Monte Carlo p-value in the continuous case,

$$p = P(T_{stat} \geq t | H_o) \approx p_{mc} = \frac{(\#t^* \geq t) + 1}{N + 1}, \text{ Davison and Hinkley [32]}$$

As in the case when dealing with the contaminant concentration data, the null distribution will depend on nuisance parameters, α, β of the Gamma model. These nuisance parameters are estimated by the MLE method to obtain the null model $\hat{F}_o$. The Monte Carlo p-value method does not exactly apply since the null distribution is dependent upon nuisance parameters with an approximate p-value of $p = P(T_{stat} \geq t | \hat{F}_o)$. The bootstrap p-value uses the same procedure except the N simulated data sets are obtained from $\hat{F}_o$, yielding the bootstrap p-value,

$$p_{boot} = \frac{\#t^* \geq t}{N}.$$

The problem of testing the lognormal model vs. the Gamma model can be formulated as,

$$H_o : f(x) = f_o(x) \text{ vs } H_a : f(x) = f_a(x)$$

where $f_o(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$, is the Gamma pdf, and $f_a(x) = \frac{e^{-1/2\sigma^2(\ln x - \mu)^2}}{\sigma x (2\pi)^{1/2}}$ is the lognormal pdf. Since the distributions contain nuisance parameters the bootstrap p-value is based on the test statistic $T_{stat} = 1/n \ln\left[\frac{L_a(x, \theta_a)}{L_o(x, \theta_o)}\right] = 1/n[\ln L_a(x, \theta_a) - \ln L_o(x, \theta_o)]$, where θ represents the ML estimates. The value of t is derived as follows,

$$1/n \ln L_a = -\frac{\sum_{i=1}^{n} (\ln x_i - \mu)^2}{2n\sigma^2} - 1/2 \ln \sigma^2 - 1/2 \ln 2\pi - \overline{\ln x}$$

which upon substituting the ML estimates, $\hat{\mu} = \overline{\ln x}$ and $\hat{\sigma}^2 = \sum (\ln x_i - \hat{\mu})^2 / n = s_{\ln}^2$,

$$= -\frac{s_{\ln}^2}{2s_{\ln}^2} - 1/2 \ln s_{\ln}^2 - 1/2 \ln 2\pi - \hat{\mu}$$

$$= -1/2 - 1/2 \ln 2\pi s_{\ln}^2 - \hat{\mu}$$

Similarly for the null hypothesis we have ,

$$1/n \ln L_o = (\alpha - 1)\overline{\ln x} - \bar{x}/\beta - \ln \Gamma(\alpha) - \alpha \ln \beta$$

$$= \alpha \overline{\ln x} - \overline{\ln x} - \bar{x}/\beta - \ln \Gamma(\alpha) - \alpha \ln \beta$$

which upon substitution of the ML estimates, $\hat{\beta} = \bar{x}/\hat{\alpha}$, where $\hat{\alpha}$ satisfies the equation

$\ln \alpha - \Psi(\alpha) = \ln \bar{x} - \overline{\ln x}$,

$$= \hat{\alpha}\hat{\mu} - \hat{\mu} - \hat{\alpha} - \ln \Gamma(\hat{\alpha}) - \hat{\alpha} \ln(\bar{x}/\hat{\alpha})$$

From these equations the value of the test statistic t from the data is,

$$t = -1/2 - 1/2 \ln 2\pi s_{\ln}^2 - \hat{\mu} - \hat{\alpha}\hat{\mu} + \hat{\mu} + \hat{\alpha} + \ln \Gamma(\hat{\alpha}) + \hat{\alpha} \ln(\bar{x}/\hat{\alpha})$$

$$= -1/2 - 1/2 \ln 2\pi s_{\ln}^2 - \hat{\alpha}\hat{\mu} + \hat{\alpha} + \ln \Gamma(\hat{\alpha}) - \hat{\alpha} \ln(\hat{\alpha}/\bar{x}),$$ Davison and Hinkley [32]

Finally random samples are generated from the fitted distribution under the null hypothesis, and for each simulated data set, t^* is calculated and ordered and the bootstrap p-value is obtained.

The following three examples are given to illustrate the above procedure. The first example is a data set generated in Minitab from a Gamma distribution with parameters $\alpha = e$ and $\beta = 1$, while the second example is a generated set from a lognormal

distribution with parameters $\mu = 0.5$ and $\sigma = 1$. The parameters were chosen such that the means were the same for both distributions. The third example was obtained from Davison and Hinkley [32] and is merely offered to verify the computations. All examples are based on $N = 1000$ bootstrap runs.

Example 1. Data: n=23, $\Gamma(\alpha = e,1)$, 1.85592, 1.38518, 5.35072, 1.56020, 2.49766, 1.74838, 0.72612, 1.87382, 1.43118, 1.35661, 6.63426,1.44249, 2.66408, 4.84421, 1.05446, 1.99104, 2.42628, 5.92583, 1.73890, 0.75220, 3.07631, 3.17012, 1.07993
The ML estimates are $\hat{\alpha} = 2.80071353$, $\hat{\beta} = 0.878439$, $\hat{\mu} = 0.71122824$, $s_{\ln} = 0.61472368$ and test statistic $t = 0.5617510750066912$
the ordered t^* are 0.4689534937, 0.74046359437, ..., 1.859562263077322, therefore substituting into $p_{boot} = \frac{\#t^* \geq t}{N}$, gives $p_{boot} = .999$, with the conclusion that there is no evidence to change from a Gamma to a lognormal model..

Example 2. Data: n=23, $LN(0.5,1)$, 1.90063, 2.70740, 6.01415, 0.70032, 1.06589. 5.54605, 0.40704, 1.85950, 2.29373, 3.31397, 0.29338, 1.47165, 0.78115, 0.78115, 0.35829, 2.22856, 2.72994, 3.30368, 4.81147, 2.83486, 1.09938, 1.93110, 1.70127, 2.21762
The ML estimates are $\hat{\alpha} = 1.951001$, $\hat{\beta} = 1.1492657$, $\hat{\mu} = 0.5292290$, $s_{\ln} = 0.83424471$ and test statistic $t = 0.1569652227$
the ordered t^* are 0.1751336, 0.243724994, ..., 1.1261359482, therefore substituting into $p_{boot} = \frac{\#t^* \geq t}{N}$, gives $p_{boot} = .967$, with the conclusion that there is no evidence to change from a Gamma to a lognormal model..

Example 3. Data: n=12, 3.0, 5.0, 7.0, 18.0, 43.0, 85.0, 91.0, 98.0, 100.0, 130.0,230.0, 487.0
The ML estimates are $\hat{\alpha} = 0.720553$, $\hat{\beta} = 93.90934$, $\hat{\mu} = 3.3706617$, $s_{\ln} = 1.64617809$ and test statistic $t = -0.52765610905$
the ordered t^* are -1.9250548289,-1.8941132408, ..., 0.7695107202 therefore substituting into $p_{boot} = \frac{\#t^* \geq t}{N}$ gives $p_{boot} = .5679$, with the conclusion that there is no evidence to change from a Gamma to a lognormal model.

Results given in Davison and Hinkley [32] for Example 3 are as follows:

The ML estimates are $\hat{\alpha}=0.707$, $\hat{\mu}=3.829$, $s_{\ln}=1.52937$ and test statistic $t=$-0.465 with $p_{boot}=0.62$. The answers of course are not exactly the same due to simulation of data sets.

The BCA percentile method assumes a parametric model with an unknown parameter θ that has an ML estimate $t=\hat{\theta}$, for the Gamma distribution this would be $t=\bar{x}$. Moreover, there exists for some unknown transformation $h(t)$, unknown skewness correction factor a and unknown bias correction factor w, a transformed estimator $U=h(T)$ for $\phi=h(\theta)$ that follows a $N(\phi-w\sigma_{\phi},\sigma_{\phi}^2)$. The approach is to find the confidence limits for ϕ and transform the results back for θ using the bootstrap distribution of T, Davison and Hinkley [32]. The steps from the Davison and Hinkley [32] are given below for the parametric case assuming no nuisance parameters.

First assume that the skewness and bias correction factors are known and define the random variable U as,

$$U=\phi+(1+a\phi)(Z-w) \text{ where } Z\sim N(0,1)$$

then,

$$\ln(1+aU)=\ln(1+a\phi)+\ln(1+a(Z-W))$$

Substituting u and the ρ percentile of the standard normal, yields the confidence limit

$$\hat{\phi}_{\rho}=u+\sigma_u\left(\frac{w+z_{\rho}}{1-a(w+z_{\rho})}\right)$$

Since the inverse function h is not known, $\hat{\theta}_\rho = h^{-1}(\hat{\phi}_\rho)$ is estimated by the resampling distribution function of T^*, denoted $\hat{K}(\hat{\theta}_\rho)$. The distribution function can be written as follows,

$$\hat{K}(\hat{\theta}_\rho) = P^*(T^* < \hat{\theta}_\rho | t) = P^*(U^* < \hat{\phi}_\rho | u) = \Phi\left(\frac{\hat{\phi}_\rho - u}{\sigma_u} + w\right) = \Phi\left(\frac{w + z_\rho}{1 - a(w + z_\rho)} + w\right)$$

The right hand side is known, therefore the confidence limit for θ is,

$$\hat{\theta}_\rho = \hat{K}^{-1}\left[\Phi\left(\frac{w + z_\rho}{1 - a(w + z_\rho)} + w\right)\right]$$

N bootstrapped estimates of t^* are generated as before, with the confidence interval defined,

$$\hat{\theta}_\rho = t^*_{N\tilde{p}} \text{ where } \tilde{p} = \Phi\left(\frac{w + z_\rho}{1 - a(w + z_\rho)} + w\right)$$

The skewness and bias correction factors are not known but are easily estimated as follows.

Bias correction factor $w = P^*(T^* < t | t) = P^*(U^* < | u) = P(U < \phi | \phi) = \Phi(w)$

$\Rightarrow w = \Phi^{-1}(\hat{K}(t))$, which is simulated by, $w = \Phi^{-1}\left(\frac{\#t^*_n \le t}{N}\right)$

and,

Skewness correction factor $a = 1/6\left(\frac{E^*(l'^*(\hat{\theta})^3)}{Var^*(l'^*(\hat{\theta}))^{3/2}}\right)$, where $l'^*(\hat{\theta})$ is the derivative of the log likelihood function of a set of data simulated from the fitted model, Davison and Hinkley [32].

Modeling site data with the Gamma distribution introduces the parameters α and β. The bias correction factor and confidence limits are calculated as above with $t^* = \bar{x}$. The skewness correction factor is based upon the least-favorable family and defined by Davison and Hinkley [32] as,

$$a = 1/6\left(\frac{E^*(l'^*_{LF}(0)^3}{\mathrm{var}^*(l'^*_{LF}(0)^{2/3})}\right)$$

The least-favorable family of the original Gamma family is obtained by holding α constant at $\hat{\alpha}$ giving $l'^*_{LF}(0) \propto \bar{y}^* - \bar{y}$. Using the ML estimates the skewness correction is defined,

$a = 1/3(n\hat{\alpha})^{-1/2}$, Davison and Hinkley [32].

Results

Example 1. Data obtained from the Dissertation by Schneider [10].

The data is the amount of Strontium found in clamshells. This example is offered for verification of my computations. Of course the numbers will not be exactly the same for many reasons including my use of double precision for all real numbers.

Data: Data size, n = 207, min= 590, max = 1438

	Schneider [10] results	My results	Δ
$\hat{\alpha}$	39.074	39.6728	0.59879
$\hat{\beta}$	25.241	24.8886	0.35238
$\alpha_{lower}^{0.05}$	31.785	32.2576	0.47256
$\alpha_{upper}^{0.05}$	46.694	47.4304	0.73640
$\hat{\alpha}$adj	38.514	39.1010	0.58704
$D_n(\hat{\alpha},\hat{\beta})$	0.0688	0.06955	0.00075
m	0.013	0.0127	0.00034

Table 2. 1 Results Example 1

Alpha is adjusted using the equation, $\hat{\alpha}\text{adj} = \hat{\alpha}(1-3/n)+2/3n$, as mentioned before because $n \geq 4$ and $\hat{\alpha} \geq 1.0$.

In testing whether the Gamma distribution with the above parameters is a good fit the smoothing function with constants obtained from Schneider [10] are used. The smoothing values for ρ=.20,.15,.10,.05,.01 are as follows,

	n=4(1)9			n=10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.055310	0.05532646	0.05529767	0.051639864	0.05166745	0.05161825
0.15	0.059167	0.05919136	0.05914786	0.054122140	0.05415375	0.05409738
0.1	0.060902	0.06093254	0.06087814	0.057705493	0.05774000	0.05767846
0.05	0.065534	0.06556924	0.06550695	0.063410336	0.06344889	0.06338014
0.01	0.081425	0.08147469	0.08138529	0.073579342	0.07363525	0.07353555

Table 2. 2 KS Critical Values for Example 1

The critical values range from 0.063 to 0.065 at ρ=0.05 and from 0.073 to 0.081 at ρ=0.01. Since $D_n(\hat{\alpha},\hat{\beta}) = 0.06955$ we would reject H_o at $\rho \geq 0.05$ and accept H_o at $\rho \leq 0.01$. However there is a significance value between 0.05 and 0.01 in which H_o would be accepted. In this case the decision on whether the Gamma distribution fits the data adequately is dependant upon how conservative of a test is desired.

Example 2. Data from the Elrama School Superfund Site in Washington County, P.A.

The data was collected from two waste piles where 26 contaminates (10 organic, 12 semi-volatile and 4 volatile compounds) were detected in both piles. A two sample nonparametric Kolmogrov-Smirnov statistic test showed that there was no difference between the distributions of contaminants between the two piles. As a result the data was combined for statistical analysis. Contaminants of concern were Toluene and Aluminum:

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. $n = 23$, max = 280000.0

The results from the raw data are:
$\bar{x} = 16478.32609$, $s_x = 58510.77518$, $s_{x(mle)} = 57224.66703$
Shapiro-Wilks Statistic = 0.313, critical value at $\rho = 0.10$ is 0.928

The results from the log transformed data are:
$\bar{y} = 4.6510016$, $s_y = 3.65947912$, $s_{y(mle)} = 3.57904118$, $cv_y = .786815272$, $H_{.95} = 7.01662$
Shapiro-Wilks statistic = 0.818

The results for the Gamma distribution are:
Shape parameter $\hat{\alpha} = 0.2532224786$, Scale parameter $\hat{\beta} = 65074.499599$,
$\alpha_{lower}^{0.95} = 0.117216162$, $\alpha_{upper}^{0.95} = 0.2626961$

The results for the lognormal distribution are:
W/ MVUE estimates: $\mu_1 = 22178.222$, $\sigma_1 = 399541.779$, $\sigma_{\mu_1} = 19513.073$,
and median = 77.9306
W/ MLE estimates : $\mu_1 = 84702.0235$, $\sigma_1 = 683530421.79661$, $cv = 809.0760821546$

Goodness OF Fitresults:
Bootstrap p-value= 0.02
K-S Statistic $D_n(\hat{\alpha},\hat{\beta}) = 0.4182902562$

	n=4(1)9			n=10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.159792	0.16232745	0.15969496	0.158503937	0.16313698	0.15832715
0.15	0.171221	0.17507098	0.17107325	0.166893781	0.17223453	0.16669028
0.1	0.180095	0.18502758	0.17990701	0.177292069	0.18310399	0.1770707
0.05	0.195037	0.20072515	0.19482032	0.193631534	0.20008554	0.19338578
0.01	0.234899	0.24283073	0.23459727	0.227467039	0.23698122	0.2271063

Table 2. 3 KS Critical Values for Toluene Example 2

The range of the critical values is from 0.193 to .200 at the 0.05 significance level and 0.22 to 0.242 at the 0.01 significance level. With a K-S statistic of 0.4182902562 it appears that the Gamma distribution $F_o(x,\hat{\alpha},\hat{\beta})$ does not provide an adequate fit for the data on the Toluene concentrations found at the Elrama Superfund Site.

Aluminum: 31900, 8030, 12200, 11300, 4770, 5730, 5410, 8420, 8200, 9010, 8600, 9490, 9530, 7460, 7700, 13700, 30100, 7030, 2730, 5820, 8780, 360, 7050, $n = 23$, max = 31900

The results from the raw data are:
$\bar{x} = 9709.565217$, $s_x = 7310.01957$, $s_{x(mle)} = 7149.340181$
Shapiro-Wilks Statistic = 0.707, critical value at $\rho = 0.10$ is 0.928

The results from the log transformed data are:
$\bar{y} = 8.927329$, $s_y = 0.845015$, $s_{y(mle)} = 0.8264409$, $cv_y = 9.465484$ E-02, $H_{.95} = 2.476592$
Shapiro-Wilks Statistic = 0.781

The results for the Gamma distribution are:

Shape parameter $\hat{\alpha}$=2.1241691, Scale parameter $\hat{\beta}$=4570.99431,
Bootstrap p-value= 0.999
$\alpha_{lower}^{0.95}$=1.091675334, $\alpha_{upper}^{0.95}$ = 3.3189492

The results for the lognormal distribution are:
W/ MVUE estimates: μ_1=10561.2573, σ_1=10071.81066, σ_{μ_1}=2041.82660,
median =7418.764
W/ MLE estimates : μ_1=10768.2242, σ_1=10993.33070, cv=1.020904698
Goodness OF Fitresults:
Bootstrap p-value= 0.999
K-S Statistics $D_n(\hat{\alpha},\hat{\beta})$= 0.19534366572

	n=4(1)9			n=10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.156289	0.15704101	0.15596397	0.152200808	0.15354432	0.15162232
0.15	0.165941	0.16707039	0.16545385	0.159651273	0.1611927	0.15898796
0.1	0.173371	0.17480557	0.17275277	0.169417776	0.17109296	0.16869702
0.05	0.187299	0.18894823	0.18658839	0.184892906	0.18675142	0.18409337
0.01	0.224165	0.22644779	0.22318374	0.214711046	0.21741161	0.21355134

Table 2. 4 KS Critical Values for Aluminum Example 2

Ranges of critical values for Aluminum are from 0.184 to 0.188 and 0.213 to 0.226 at the 0.05 and 0.01 significance levels. With a K-S statistic of 0.1953 we would accept H_o at the ρ=0.01 level and just up to the ρ=0.05 level. So it appears that the Gamma distribution with the above stated parameters appears to be a good fit to the data.

Parametric results for 95% UCL are:

	Toluene	Aluminum
Central Limit Theorem	36546.66536	12216.79491
Adjusted Central LimitTheorem	47314.49136	12894.88030
Chebychev	71013.84590	16522.93790
Gamma UCL	40952.74910	12516.47299

Gamma UCL Corrected	44877.26388	12838.91109
lognormal w/ ml estimates	43876.85288	30251.31360
lognormal w/ H statistic	20201965.48	16823.50414
Chebychev Ln w/ mvue estimates	109401.6598	19688.22224
BCA percentile method	35196.3783	11260.598156

Non-parametric bootstrapping results for 95% UCL are:

	Toluene	Aluminum
Bias estimate Br	-15.09932609	-25.9769565217
Var estimate Vr	1208.654460	1547.76356500
Basic	31103.23913	12026.9565217
Studentized w/ vl	165001.5877	15486.3444570
Standard w/ Normal estimate	36361.60414	12281.4584619
Standard w/ Normal vl estimate	36105.54909	12161.6842309
Basic Percentile	39939.67391	12462.1739130
Bca percentile	54734.77788	13477.1587382

Q-Q Plots of T^*

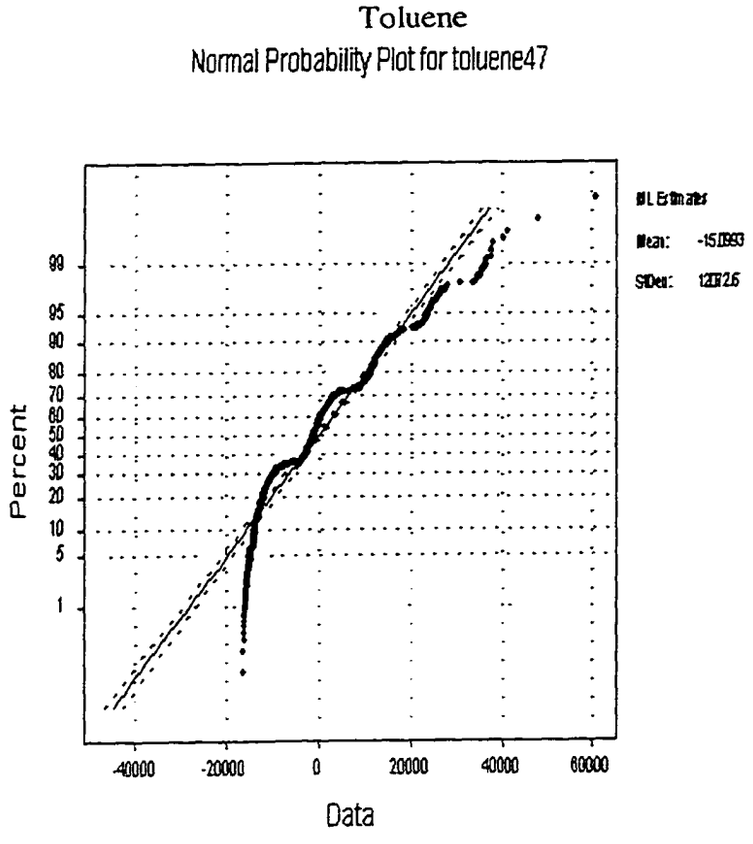

Figure 2. 1 Q-Q Plot for Toluene Example 2

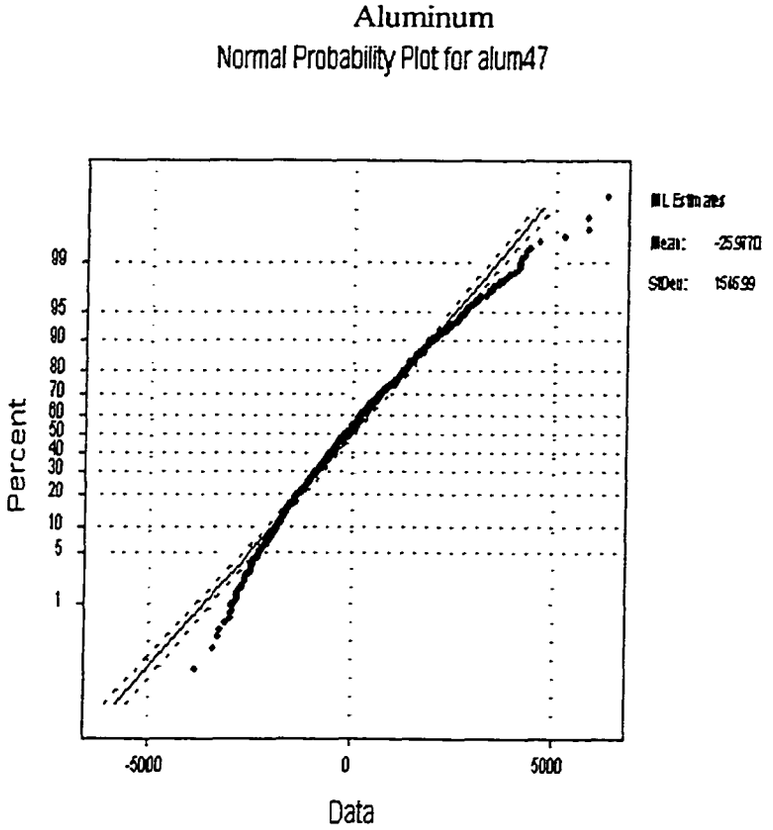

Figure 2. 2 Q-Q Plot for Aluminum Example 2

The $Q-Q$ plots of t^* for the Toluene data indicate that improvement in the non-parametric bootstrapped estimates will be gained by not using quantiles of the normal and studentized distributions. This is seen in the basic and BCA percentile limits being closer to the Gamma and lognormal MLE upper bounds. For the Aluminum data t^* appears to have departures from normality in both tails at the .05 and 0.95 levels respectively, but not as large an improvement is gained by not assuming normality. This is observed with all of the estimates falling in a similar range.

Discussion Example 2

Toluene:

The first and most obvious observation of the results obtained for the Toluene data is the H-statistic based UCL being orders of magnitude larger than those obtained from other methods. Moreover this UCL is even orders of magnitude larger than the maximum data point and both conservative Chebychev UCL's. An explanation for this extreme value is the sensitivity of the H-based statistic when the lognormal distribution is highly skewed, as is the case with an observed standard deviation of 3.5659 in the log transformed data. The lognormal distribution also failed for lack of fit with a Shapiro-Wilks statistic of 0.818 for the log transformed data with a critical value of 0.928 at the 0.10 significance level.

The Gamma UCL's are in the middle of all the parametric and nonparametric bootstrapped intervals and are almost the same as the lognormal UCL, using the ML estimates. Although the Gamma distribution failed for Lack of Fit it passed verse an alternative lognormal distribution with a bootstrap p-value of 0.99. Therefore the Gamma and lognormal with ML estimates produce more appropriate bounds in comparison with parametric and most importantly nonparametric bootstrapped bounds.

Aluminum:

Observe that all of the bounds fall in the range from 12,026 to 16,522 with the exception of the bounds using an assumed lognormal distribution. The worst bound is the ML estimated lognormal UCL, which has a difference of more than 10,000 compared to both Chebychev bounds. The H-based UCL is significantly higher than the other bounds and lies in the same region as the Chebychev bounds. The Shapiro-Wilks statistic is

0.781, thus again a lognormal distribution fails for lack of fit at the 0.10 significance level. Note however if the CV test is used, which is suggested by the EPA and many environmental scientists, a CV value of 0.75 for the raw data would erroneously lead to assuming a normal distribution despite the strong rejection of the Shapiro-Wilks test with a p-value of 0.00002 at the 0.10 significance level.

The Gamma distribution however passes the lack of fit test right up to the 0.05 significance level and verse an alternative lognormal distribution, with a bootstrap p-value of 0.99. More evidence for the appropriateness of a Gamma modeled UCL is seen in its closeness to all the parametric and nonparametric bounds.

Example 3. Data from the Naval Construction Battle Center (NCBC) Superfund Site in Rhode Island

Analysis was performed on inorganic compounds in groundwater from 17 wells at the NCBC for the purpose of providing reliable mean background threshold levels for the various contaminates at the site. Results for two contaminates, Aluminum and Manganese are provided below.

Aluminum: 290, 113, 264, 2660, 586, 71, 527, 163, 107, 71, 5920, 979, 2640, 164, 3560, 132, 125, $n = 17$, max = 5920

The results from the raw data are:
$\bar{x} = 1849.41176470$, $s_x = 3351.272576$, $s_{x(mle)} = 3251.2119560$

The results from the log transformed data are:
$\bar{y} = 6.2256806317$, $s_y = 1.6592604430$, $s_{y(mle)} = 1.6097190746$, $cv_y = 0.266518721$, $H_{.95} = 4.369302$

The results for the Gamma distribution are:
Shape parameter $\hat{\alpha} = 0.50870$, Scale parameter $\hat{\beta} = 3635.4956$, Bootstrap p-value= 0.081
$\alpha_{lower}^{0.95} = 0.262116084$, $\alpha_{upper}^{0.95} = 0.7986834$

The results for the lognormal distribution are:
W/ MVUE estimates: $\mu_1 = 1718.046346$, $\sigma_1 = 4062.004$, $\sigma_{\mu_1} = 810.72973351$,
median =465.9477
W/ MLE estimates : $\mu_1 = 2002.70338$, $\sigma_1 = 7676.367263$, $cv = 3.83300258$

Goodness OF Fitresults:
Bootstrap p-value= 0.081
K-S Statistics $D_n(\hat{\alpha}, \hat{\beta}) = 0.22595584032$

	n =4(1)9			n =10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.181104	0.18279463	0.18024146	0.179615509	0.18269669	0.17805307
0.15	0.192791	0.19533889	0.19149535	0.188776129	0.19231949	0.18698148
0.1	0.202732	0.20599208	0.20107819	0.200201558	0.20405045	0.19825282
0.05	0.219464	0.22322037	0.21755946	0.218216667	0.22248178	0.2160577
0.01	0.261271	0.26644399	0.2586534	0.254817144	0.26105645	0.25167042

Table 2. 5 KS Critical Values for Aluminum Example 3

The ranges of the critical values for the rejection of H_o are from 0.2176 to 0.22 and 0.251 to 0.266 at the 0.05 and 0.01 significance levels. With a value of 0.225 for the test statistic, H_o is accepted at $\rho = 0.01$. Since the test statistic is equal to some of the critical values up to the second decimal place and only 0.008395 greater than the smallest critical value H_o is accepted at $\rho < 0.05$. The Gamma distribution with the above parameters does provide a reasonably good fit to the Aluminum concentrations found in the groundwater at the Rhode Island Superfund Site.

Manganese: 15.8, 28.2, 90.6, 1490, 85.6, 281, 4300, 199, 838, 777, 824, 1010, 1350, 390, 150, 3250, 259, $n = 17$, max = 4300

The results from the raw data are:
$\bar{x} = 902.247058$, $s_x = 1189.488513$, $s_{x(mle)} = 1153.973359$

The results from the log transformed data are:
$\bar{y} = 5.9121327$, $s_y = 1.56766586$, $s_{y(mle)} = 1.5208592$, $cv_y = 0.2651608048$, $H_{.95} = 4.097385$

The results for the Gamma distribution are:
Shape parameter $\hat{\alpha} = 0.687085205$, Scale parameter $\hat{\beta} = 1313.15163$,
Bootstrap p-value= 0.397
$\alpha_{lower}^{0.95} = 0.344235958$, $\alpha_{upper}^{0.95} = 1.1024264$

The results for the lognormal distribution are:
W/ MVUE estimates: $\mu_1 = 1108.27114$, $\sigma_1 = 2394.03912$, $\sigma_{\mu_1} = 491.514943$, median =343.54586
W/ MLE estimates : $\mu_1 = 1262.5905$, $\sigma_1 = 4125.49911$, $cv = 3.2674878$

Goodness OF Fitresults:
Bootstrap p-value= 0.397
K-S Statistics $D_n(\hat{\alpha}, \hat{\beta}) = 0.11926291037$

	n =4(1)9			n =10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.181104	0.18279463	0.18024146	0.179615509	0.18269669	0.17805307
0.15	0.192791	0.19533889	0.19149535	0.188776129	0.19231949	0.18698148
0.1	0.202732	0.20599208	0.20107819	0.200201558	0.20405045	0.19825282
0.05	0.219464	0.22322037	0.21755946	0.218216667	0.22248178	0.2160577
0.01	0.261271	0.26644399	0.2586534	0.254817144	0.26105645	0.25167042

Table 2. 6 KS Critical Values for Manganese Example 3

The ranges of the critical values for Manganese found in the groundwater are from 0.251 to 0.266 at $\rho = 0.01$ and 0.178 to 0.1827 at $\rho = 0.2$. The test statistic value of 0.1192 suggests the Gamma distribution with the above parameters clearly provides a good fit with a significance level greater than 0.2.

Parametric results for 95% UCL are:

	Aluminum	Manganese
Central Limit Theorem	3186.391401	1376.78975
Adjusted Central Limit Theorem	3675.804133	1503.796125
Chebychev	5482.641125	2191.812288
Gamma UCL	3707.415358	1603.813874
Gamma UCL Corrected	3971.838398	1712.085861
lognormal w/ ml estimates	7746.513378	4869.693602
lognormal w/ H statistic	12267.46278	6290.233486
Chebychev Ln w/ mvue estimates	5342.008255	3305.3429420
BCA percentile method	3683.944376	1581.163990

Non-parametric bootstrapping results for 95% UCL are:

	Aluminum	Manganese
Bias estimate Br	3.14682352941	2.88354117646
Var estimate Vr	800.87406686	283.118380784
Basic	2991.88235294	1329.54117647
Studentized w/ vl	5701.16823894	1993.35781027
Standard w/ Normal estimate	3163.62269375	1365.06494219
Standard w/ Normal vl estimate	3146.47253204	1362.62109531
Basic percentile	3325.17647058	1405.84705882
Bca percentile	3878.49841450	1523.40715230

Q-Q Plots of T^*

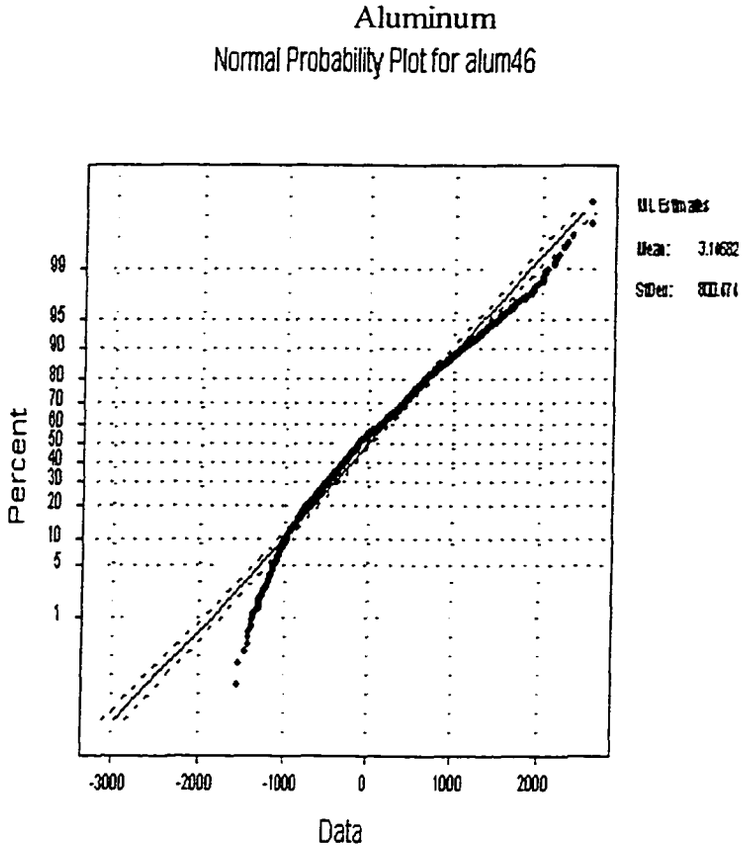

Figure 2. 3 Q-Q Plot for Aluminum Example 3

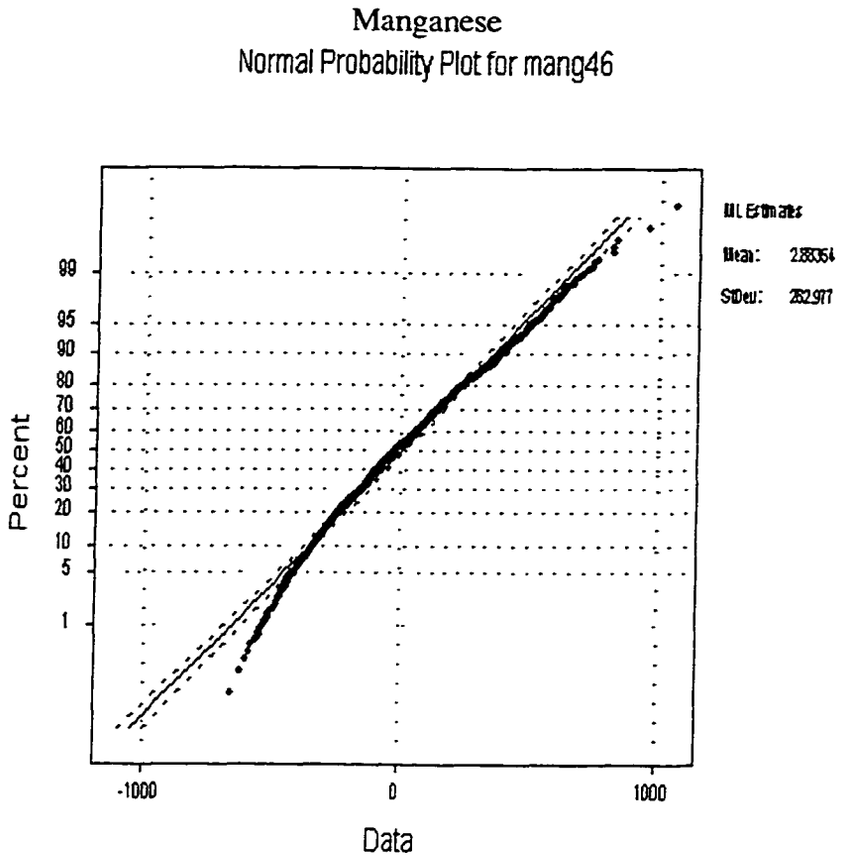

Figure 2. 4 Q-Q Plot for Manganese Example 3

The $Q-Q$ plots of t^* for the Aluminum data shows a marked departure from normality in the bottom tail and a less significant difference in the upper tail which suggests nonparametric bootstrap intervals assuming normality may be effected. The middle and upper tails of the manganese $Q-Q$ plot of t^* follow a normal distribution indicating that the above nonparametric upper bounds involving a normal assumption are justified. This is observed with all of the nonparametric bounds falling in a similar range.

Discussion Example 3

Aluminum:

For the Aluminum data the Shapiro-Wilks lower tailed test statistic is 0.913 with a critical value of 0.91 at the 0.01 significance level, which implies that a lognormal distribution just passes and can be assumed to model the data. The lognormal MLE based UCL of 7,746.51 is comparable with the other methods, with the latter mentioned estimates tending to be conservative 95% confidence bounds based on a lognormal assumption with values ranging between 2,991 to the Chebychev bound of 5,482. However, we again see an H-based UCL significantly larger than the MLE based UCL associated with a log transformed standard deviation of 1.659.

The Gamma distribution not only passes the lack of fit test at the 0.01 significance but also is only off in the third decimal place from passing at the 0.05 level. This implies that the Gamma distribution may indeed provide a better fit to the data considering the influence in the mean associated with skewness in the lognormal distribution. This is evidenced by the value of the Gamma UCL being consistent with the nonparametric methods, Adjusted Central Limit, BCA percentile UCL's, and smaller than the conservative Chebychev UCL as expected. Moreover the bootstrap p-value for testing the Gamma distribution verses an alternative lognormal distribution is 0.081 suggesting a Gamma model to be an acceptable alternative.

Manganese:

The log transformed Manganese data results show a medium level of skewness in an assumed lognormal model with a standard deviation of 1.567. The H-based bound is larger than the MLE based estimate and doubles the lognormal based Chebychev UCL.

However, the Shapiro-Wilks statistic for the log transformed data is 0.725, which implies that a lognormal distribution is appropriate at the 0.01 significance level.

Observe that the Gamma distribution has a bootstrap p-value of 0.397 verse an alternative lognormal distribution and passes the lack of fit test with a significance level greater than 0.2. This by itself is strong evidence for a Gamma model providing a better fit to the data and is further strengthened by noting that it's value of 1712.085 is consistent with all the other bounds and less than the Chebychev bound.

Example 4.

The data was generated from a lognormal distribution with a mean of 5 and standard deviation of 1.5. $LN(5,1.5)$

Data: 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,
$n = 15$, max = 1857.7698

The results from the raw data are:
$\bar{x} = 631.645026$, $s_x = 603.13363$, $s_{x(mle)} = 582.682444$

The results from the log transformed data are:
$\bar{y} = 5.7605$, $s_y = 1.5364482$, $s_{y(mle)} = 1.484349$, $cv_y = 0.26672129$, $H_{.95} = 3.771535$

The results for the Gamma distribution are:
Shape parameter $\hat{\alpha} = 0.859392$, Scale parameter $\hat{\beta} = 734.990517$,
Bootstrap p-value= 0.729
$\alpha_{lower}^{0.95} = 0.395417543$, $\alpha_{upper}^{0.95} = 1.4238142$

The results for the lognormal distribution are:
W/ MVUE estimates: $\mu_1 = 901.786213$, $\sigma_1 = 1831.7585$, $\sigma_{\mu_1} = 407.578914$, median = 293.26849
W/ MLE estimates : $\mu_1 = 1033.63514$, $\sigma_1 = 3202.281495$, $cv = 3.0980772$

MLE estimated 80th, 90th and 95th percentiles of the lognormal distribution:

MLE 95^{th} percentile = 1750.00
90th percentile = 1010.00
80th percentile = 524.492

Goodness OF Fitresults:
Bootstrap p-value= 0.729
K-S Statistics $D_n(\hat{\alpha}, \hat{\beta}) = 0.0634335288$

	n=4(1)9			n=10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.190354	0.19200256	0.18958289	0.188621017	0.19162239	0.1872254
0.15	0.201958	0.20443435	0.20080395	0.198042923	0.20149086	0.1964414
0.1	0.212282	0.21544848	0.21080903	0.209859322	0.21360147	0.20812167
0.05	0.229735	0.2333824	0.22803919	0.228529063	0.23267196	0.22660574
0.01	0.272003	0.27699808	0.26968488	0.266000058	0.27204	0.26320529

Table 2. 7 KS Critical Values for Example 4

The test statistic value of 0.0634 clearly suggests that the Gamma distribution with the above parameters provides a good fit to the data obtained from a LN(5,1.5) distribution.

Parametric results for 95% UCL are:

Central Limit Theorem	887.802727526
Adjusted Central Limit Theorem	919.787992130
Chebychev	1327.75112603
Gamma UCL	1085.65981107
Gamma UCL Corrected	1193.9064470
lognormal w/ ml estimates	3975.09625778
lognormal w/ H statistic	4863.69380900
Chebychev Ln w/ mvue estimates	2723.66396077
BCA percentile method	1078.57956128

Non-parametric bootstrapping results for 95% UCL are:

Bias estimate Br	7.63381789333
Var estimate Vr	149.244287172
Basic	631.645026666
Studentized w/ vl	946.046377370
Standard w/ Normal estimate	869.503136743
Standard w/ Normal vl estimate	879.116876640
Basic percentile	903.244420000
Bca percentile	926.259540923

Graphs Example 4

The first graph is a comparison of the bounds assuming a Gamma verse a lognormal model. The second graph plots the results of the Gamma distribution with the BCA percentile bootstrapped estimate, and the Chebychev and Central Limit Theorem based bounds.

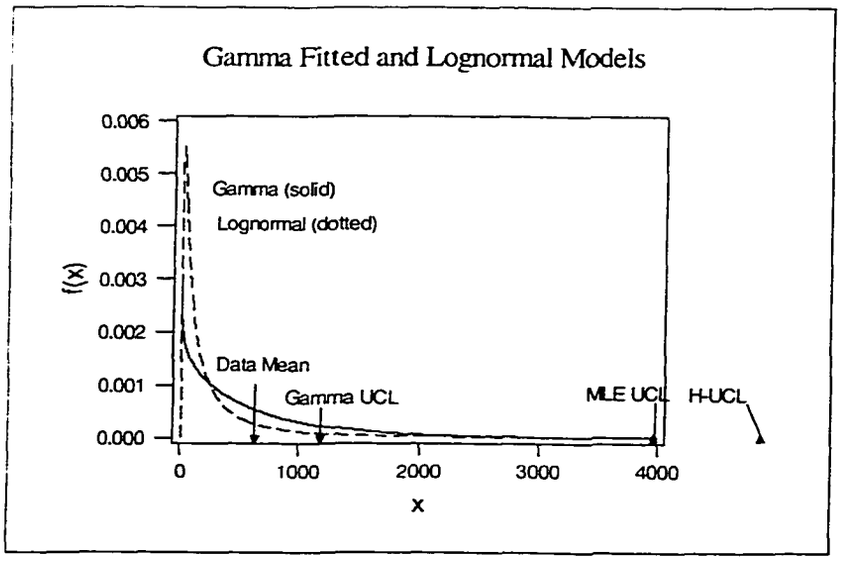

Figure 2. 5 Plot of Gamma and Lognormal Models Example 4

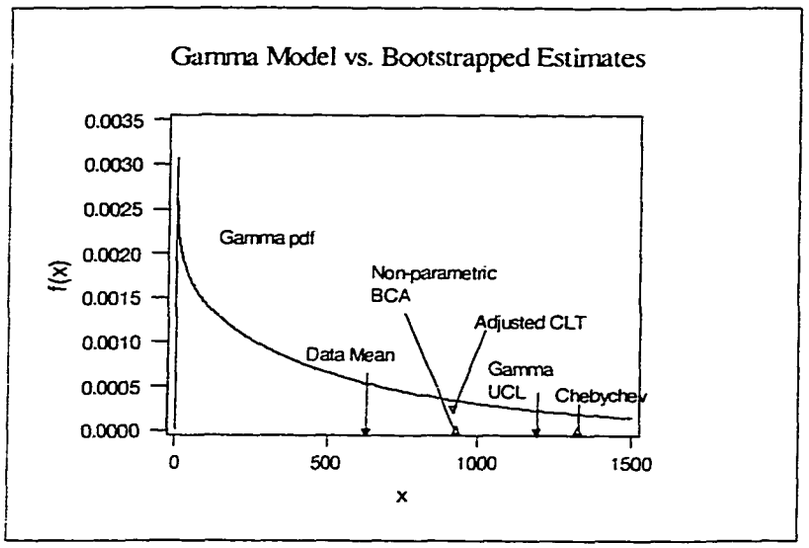

Figure 2. 6 Plot of Gamma and Bootstrapped Estimates Example 4

Discussion Example 4

The 80th, 90th and 95th MLE estimated percentiles for the lognormal distribution are 524.4929, 1010.0 and 1750.0, respectively. Note the H-based UCL is considerably higher than the 95th percentile MLE estimate, with a value of 4863.6938. All bootstrapping estimates are conservative with values near the 80th percentile, which is expected since many are based on normal assumptions.

The Chebychev and Gamma UCL corrected bounds are greater than the 80th and less than the 90th as would be expected when constructing a 95% confidence bound for the mean. Moreover, the Gamma distribution clearly passed the lack of fit test and is accepted verse an alternative lognormal distribution.

Example 5.

The data was generated from a lognormal distribution with a mean of 5 and standard deviation of 1.7. $LN(5,1.7)$

Data: 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, $n=15$, max = 1860.4443

The results from the raw data are:
$\bar{x}=473.413113$, $s_x=606.79239$, $s_{x(mle)}=586.2171429$

The results from the log transformed data are:
$\bar{y}=5.17839050$, $s_y=1.7100675$, $s_{y(mle)}=1.65208223$, $cv_y=0.33023$, $H_{.95}=3.990634$

The results for the Gamma distribution are:
Shape parameter $\hat{\alpha}=0.634825$, Scale parameter $\hat{\beta}=745.73787$, Bootstrap p-value=0.459

$\alpha_{lower}^{0.95}=0.303215588$, $\alpha_{upper}^{0.95}=1.0387697$

The results for the lognormal distribution are:
W/ MVUE estimates: $\mu_1=636.34201$, $\sigma_1=1521.6358$, $\sigma_{\mu_1}=322.05291$, median=160.7589
W/ ML estimates : $\mu_1=765.5205$, $\sigma_1=3213.524178$, $cv=4.1978287$

MLE estimated 80th, 90th and 95th percentiles of the lognormal distribution:

MLE 95th percentile = 2430.00
90th percentile = 1310.00
80th percentile = 620.643

Goodness OF Fitresults:
Bootstrap p-value= 0.459
K-S Statistics $D_n(\hat{\alpha},\hat{\beta})=0.1906260701755$

	n=4(1)9			n=10(5)30		
	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)	A(ρ)
Significance Level ρ	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$	With $\hat{\alpha}$	With $\alpha_{lower}^{0.95}$	With $\alpha_{upper}^{0.95}$
0.2	0.190921	0.19273478	0.19004085	0.18965047	0.19296208	0.18805406
0.15	0.202809	0.20553668	0.20148961	0.199224972	0.20303151	0.19739218
0.1	0.213369	0.21686073	0.21168389	0.211142062	0.21527408	0.20915323
0.05	0.230986	0.23501033	0.22904619	0.229949035	0.23452408	0.22774749
0.01	0.273715	0.27923141	0.27106114	0.268067186	0.27474895	0.26486365

Table 2. 8 KS Critical Values for Example 5

The test statistic value of 0.1906 clearly suggests that the Gamma distribution with the above parameters provides a good fit to the data obtained from a LN(5,1.7) distribution.

Parametric results for 95% UCL are:

Central Limit Theorem	731.124731140
Adjusted Central Limit Theorem	781.243836956
Chebychev	1173.74196712
Gamma UCL	910.206821523
Gamma UCL Corrected	966.687286900
lognormal w/ ml estimates	2955.07899276
lognormal w/ H statistic	4742.94910000
Chebychev Ln w/ mvue estimates	2075.91855325
BCA percentile method	897.784371234

Non-parametric bootstrapping results for 95% UCL are:

Bias estimate Br	7.54915368666
Var estimate Vr	149.753442566
Basic	473.413113333
Studentized w/ vl	823.505691297
Standard w/ Normal estimate	712.193397323
Standard w/ Normal vl estimate	722.386189701
Basic percentile	747.344653333
Bca percentile	794.732123276

Graphs Example 5

The following are two graphs of the results of example 5, comparing the results assuming a Gamma model to the lognormal and selected non-parametric bootstrapped estimates.

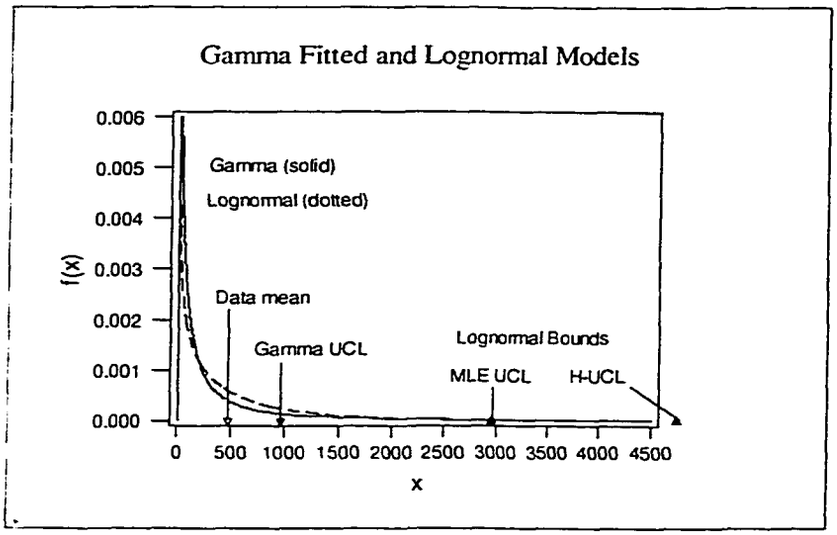

Figure 2. 7 Plot of Gamma and Lognormal Models Example 5

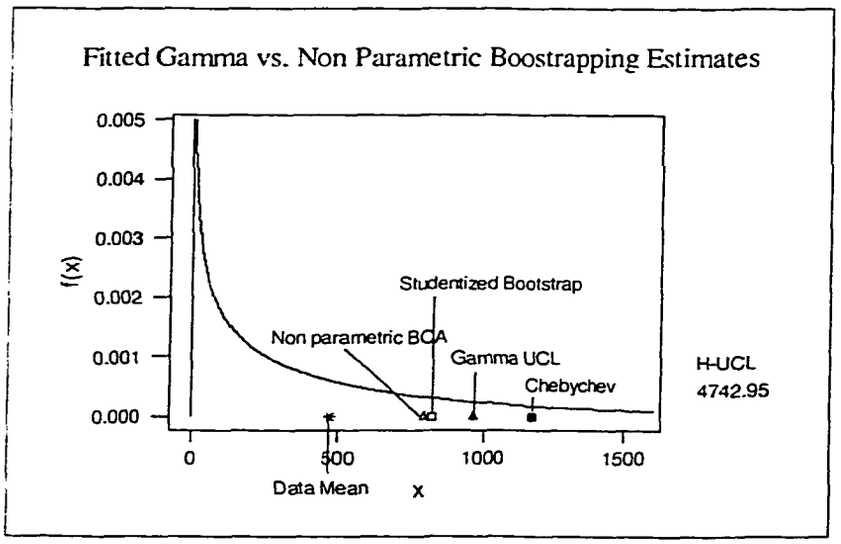

Figure 2. 8 Plot of Gamma and Bootstrapped Estimates Example 5

Discussion Example 5

The 80th, 90th and 95th percentiles for the lognormal distribution are 620.643, 1310.0 and 2430.0, respectively. Note the H-based UCL is considerably higher than the 95th

percentile MLE estimate with a value of 4742.9491. All bootstrapping estimates are conservative with values near the 80^{th} percentile, which is expected since many are based on normal assumptions.

The Chebychev and Gamma UCL corrected bounds are greater than the 80^{th} and less than the 90^{th} as would be expected when constructing a 95% confidence bound for the mean. Moreover, the Gamma distribution clearly passed the lack of fit test and is accepted verse an alternative lognormal distribution.

Example 6.

The purpose of the following example is to see how the H, MLE, and Gamma corrected 95% UCL's react to varying levels of sample size and skewness. Sample sizes of 10, 20 and 30 were generated from a lognormal distribution with a constant mean of 5 for each values, σ = 0.5, 1.0, 1.5, 2.0, and 3.0. The 80^{th}, 90^{th}, and 95^{th} percentiles of a single random variable for the specified lognormal distribution are calculated for comparative purposes. The value of the H statistic is based on the sample standard deviation. Below are the tabulated results.

	n=10			n=20			n=30		
	H-based UCL	MLE UCL	Gamma UCL	H-based UCL	MLE UCL	Gamma UCL	H-based UCL	MLE UCL	Gamma UCL
σ_1									
0.5	316.28761	424.989	282.3198	239.82255	388.21	229.5477	227.02382	379.02	226.2984
	s_x=0.64			s_x=0.557			s_x=0.495		
1.0	367.117	469.203	407.4237	522.9199	894.06	539.8672	473.961	926.97	371.8049
	s_x=0.7671			s_x=1.0292			s_x=1.076		
1.5	37146.492	4177.44	1494.822	2746.0493	2721.7	1007.2484	1092.8259	1928.4	641.6404
	s_x=1.962			s_x=1.600			s_x=1.3252		
2.0	375198.26	8135.28	3806.946	62850.399	12382	4678.8322	8071.86141	5649.4	2051.939
	s_x=2.3472			s_x=2.3910			s_x=2.1234		
3.0	2.612E+12	60287.1	10179.40	75529.055	6679.0	1893.5295	18220.433	5105.2	1005.964
	s_x=4.21739			s_x=2.6438			s_x=2.5434		

Table 2. 9 Results Example 6

σ_1	80th Percentile	90th Percentile	95th Percentile
0.5	226.06	281.68	337.79
1.0	344.33	534.61	768.81
1.5	524.49	1010.0	1750.0
2.0	798.90	1930.0	3980.0
3.0	1850.0	6940.0	20600.

Table 2. 10 Pecentiles of Lognormal Distribution Example 6

Graphs Example 6

The following are two graphs from a lognormal distribution with mean 5 and σ = 2.0 and 3.0. These graphs show the problems of using the mean of a lognormal distribution as a measure of central tendancy when the standard deviation is large. Note the means position in relation to the median.

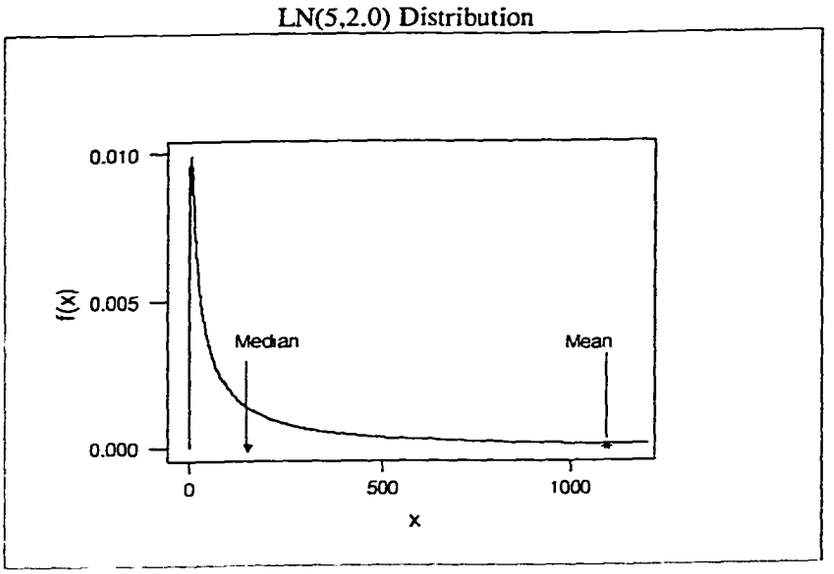

Figure 2. 9. Plot 1 of Mean and Median Example 6

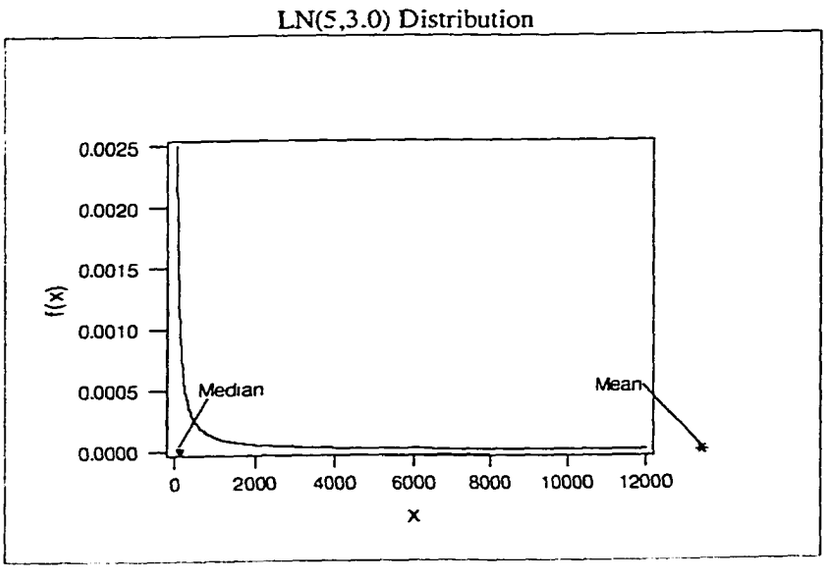

Figure 2. 10 Plot 2 of Mean and Median Example 6

Discussion Example 6

In this example for $\sigma = 0.05$ and 1.0 the H based and Gamma corrected UCL's all fall below the 95th percentiles for each sample size while the MLE based UCL is larger for every data set except, $n=10, \sigma = 1.0$. Additionally the H based upper bounds are similar to the Gamma corrected estimates with a difference of only 0.72542 when $n=30$ and $\sigma = 0.05$. As expected the Gamma distribution yields very stable estimates across these sample sizes with the entire Gamma based limits falling around the 90th percentiles.

For the data sets with $\sigma = 1.5$, both the H and MLE based upper bounds exceed the 95th percentiles for all n. Note also the case when $n = 10$, the H-UCL is an order of magnitude larger than the 95th percentile. Unlike the preceding two values of σ when n is greater than 10, the H and MLE based bounds are similar, with both exceeding the 95th percentile. Again the Gamma based bounds prove to be robust at this level of skewness in the data with all estimates falling very close to the 90th percentile.

When the skewness increases, perceptible by a standard deviation of 2.0, the H based bound starts to drastically exceed all other estimates. When the standard deviation is 2.0 we see that both the H and MLE based upper bounds exceed the 95th percentile for all values of n. When n is 10, the MLE is more than double this value while the H based upper bound is 2 orders of magnitude larger. Moreover, when the sample size is increased to 20, observe that the MLE based bound is 3 times the 95th percentile while the H based UCL is more than 10 times larger respectively. Much improvement is seen when the sample size is increased to 30, but the H-UCL is still almost double the 95th percentile. With the Gamma corrected upper bound for the mean falling just below the 95th percentile, when the sample size is 10, shows it's ability to handle a high level of

skewness coupled with a small sample size. At a sample size of 20 it slightly exceeds the 95th percentile but in considering randomness in the data and the excessively high values of the lognormal based bounds, it is quite reasonable. In addition at n=30 it falls on the 90th percentile.

Again when the standard deviation is increased to 3.0, the H based value becomes very unstable especially when the sample size is 10 with an astronomical value of 2.612E +12. Even though the H-UCL decreases to 75,529 at a sample size of 20 it's still much larger than the MLE, and Gamma corrected UCL's and the 95th percentile. At $n = 30$, the H based falls just beneath the 95th percentile and is 3 times larger than the 90th percentile. Surprisingly the MLE estimated bounds appear reasonable with exception when the sample size is 10, however this was not the case when σ was 1.5 and 2.0, reflecting the sensitive nature of the lognormal distribution. The Gamma based bounds are very consistent across the sample sizes tending to be more conservative at this level of standard deviation with values falling very close to the 80th percentiles.

Conclusions

The lognormal distribution is typically used to model contaminant concentration data from Superfund Sites to provide upper confidence limits for the mean contaminate concentration. The verification and appropriateness of its use is often based on the less robust CV test. The two Superfund Site examples showed that a CV test could lead to acceptance of a lognormal distribution when in fact the more statistically accepted Shapiro-Wilks test significantly rejects such an assumption. Additionally the use of the lognormal H-based UCL is also suggested, while it has optimal theoretical properties the practical merit of its use in environmental applications is questionable. As the CV starts

exceeding 1.0 or the sample size is decreased or any combination of the two we see an H-based UCL often orders of magnitude larger than the maximum observed data point. This behavior was even observed when data was generated from a lognormal distribution, which produced estimates orders of magnitude larger than the respective percentile for a single random variable. This was experiential in examples 4,5 and 6 where all estimates based on a lognormal distribution continuously provided inconsistent and unreasonable results emphasizing its sensitive mathematical properties. These estimates are often overstated, which can lead to over estimating the upper bound for the mean contaminate concentration. Such a situation could lead to non-cleanup of a contaminated site with the possibility of threatening the environment, humans or both.

In this study, the two-parameter Gamma distribution is proposed as an alternative to the lognormal model. This distribution provided upper bounds consistent with all of the parametric and nonparametric bounds. Moreover, it generally fell above these estimates and below the conservative Chebychev estimates. When the data was indeed from a lognormal distribution it fell between the 80^{th} and 95^{th} estimated percentiles, as would be expected, and passed the Goodness Of Fit Test. Regardless of the sample size or skewness of the Superfund data sets it consistently provided reasonable estimates of the upper bounds for the site contaminated data. Additionally, the lognormal model was only accepted twice at the 0.01 significance level in the four Superfund data sets while the Gamma distribution was accepted three out of the four times up to and greater than the 0.05 significance level. In consequence the observed results based on the Gamma distribution provide a much more stable and preferable modeling distribution of contaminant concentration data.

References

[1] EPA(1989), "Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities", Publication PB89-151047, April 1989.

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

[3] Linhart, H. Approximate Confidence Limits For The Coefficient Of Variation Of Gamma Distributions. Biometrics, Sept., 733. (1965)

[4] Gilbert. Richard O., " Statistical Methods For Environmental Pollution Monitoring," Van Nostrand Rheinhold, New York, NY, 1987.

[5] Gilbert, Richard O., "Comparing Statistical Tests for Detecting Soil Contamination Greater that Background," Pacific Northwest Laboratory, Technical Report No. DE 94-005498, 1993.

[6] Singh, Ashok K., Singh Anita, Englehardt, Max. " The Lognormal Distribution in Environmental Applications," EPA Technology Support Center Issue No. 600/R-97/006, December 1997.

[7] Johnson, N.L., Kotz, S., Continuous Univariate Distributions. Houghton-Mifflin, Boston, Mass, 1970.

[8] Grice, John V., Bain, Lee J. Inferences Concerning the Mean of the Gamma Distribution. Journal of the American Statistical Association, Dec., v. 75, No. 372, 929. 1980.

[9] Fisher, R.A., On the Mathematical Foundations of Theoretical Statistics. Trans. Royal Society London Ser. A 222, p. 309-368

[10] Schneider, Bruce E. Kolmogrov-Smirnov Test Statistics for the Gamma For the Gamma Distribution With Unknown Parameters. Dissertation, Temple University, 1978

[11] Choi, S.C., Wette, R. Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias. Technometrics, v. 11, No. 4, 683. 1969.

[12] Pairman, E., Tables of the Digamma and Trigamma Function. Cambridge University Press, Cambridge, England, 1954.

[13] Jordan, C., Calculus of Finite Differences. Chelsea Publishing Company, New York, NY, 1960.

[14] Mood, A.M., Graybill, F.A., Boes, D.C., Introduction to the Theory of Statistics McGraw Hill, New York, NY, 1974.

[15] Bartlett, M.S., Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society, A.160, 268-282, 1937.

[16] Bishop, D.J., Nair, U.S., A Note On Certain Methods of Testing for the Homogeneity and the Logarithmic Transformation. Journal of the Royal Statistical Society, Suppl. 6, 89-99, 1939.

[17] Conover, W. J. Practical Nonparametric Statistics 2ed. John Wiley & Sons, New York, NY, Chapter 6, 1980.

[18] Kolmogorov, A.N., Sulla Determinazione Empirica di Una Legge di Distibuzione. Giornale dell' Istituto Italiano degli Attuari, 4, 83-91.

[19] Smirnov, N.V., Table for Estimating Goodness Of Fit of Empirical Distributions. The Annals of Mathematical Statistics, 19, 279-281, 1948.

[20] Lee, S.W., The Power of the One-Sided and One Sample Kolmogrov-Smirnov Test. Unpublished master's report, Kansas State University, 1966.

[21] Durbin, J., Kolmogrov-Smirnov Tests When Parameters Are Estimated With Applications to Tests of Exponentiality and Tests on Spacings. Biometrika, 62, 5-22, 1975.

[22] Whittaker, J. Generating Gamma and Beta Random Variables with Non-integral Shape Parameters. Appl. Statist, 23, No. 2, 210. 1974.

[23] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes in Fortran, Cambridge University Press, 1994.

[24] Koch, G.S. Jr., Link, R.F., Statistical Analyses of Geological Data. Dover, New York , Vol 1,2, 1980.

[25] Bradu, D., Mundlak, Y., Estimation in Lognormal Linear Models. Journal of the American Statistical Society 65 (198-211), 1970.

[26] Finney, D.J., " On the Distribution of a Variate whose Logarithmic is Normally Distributed," Supplement to the Journal of the Royal Statistical Society, Vol 7, p. 155-161, 1951.

[27] Aitchison, J., Brown, J.A. C., The Log-Normal Distribution, Cambridge University Press, Cambridge, Mass, 1966.

[28] Land, C.E., Confidence Intervals for Linear Functions of the Normal Mean And Variance. Annals of Mathematical Statistics, 42, 11187-1205, 1971.

[29] Land, C.E., 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, 1975.

[30] Abramowitz, M., Stegun, I.A., Handbook of Mathematical Functions. Dover Publications, Inc., New York, 1964.

[31] Gerald, C., Wheatley, P.O., Applied Numerical Analysis. Addison-Wesley Publishing Company, Reading, Mass. 1970.

[32] Davison, A.C., Hinkley, D.V. Bootstrap Methods and their application. Cambridge University Press, New York, NY, 1997.

CHAPTER 3

DETAILS AND DERIVATIONS OF TOPICS IN CHAPTER 2

This chapter contains details of derivations of results given in Chapter 2.

Section 1: Numerical Methods for Computing the Maximum Likelihood Estimates For the Parameters of the Gamma Distribution.

A more detailed look at the computation and numerical methods for estimating the parameters of a Gamma distribution presented in Chapter 2, Section 1.

Definitions and Properties

The two-parameter Gamma model with shape parameter α and scale parameter β has the probability density function,

$$f(x,\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}dx,\ 0 \le x \le \infty,\ \alpha \ge 0, \beta \ge 0$$

with $\mu = \alpha\beta$ and $\sigma = \alpha\beta^2$.

The moment generating function is obtained in the usual way by finding the expected value of e^{tx},

$$M_x(t) = E(e^{tx}) = \int_0^\infty e^{tx} \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} dx = \int_0^\infty \frac{x^{\alpha-1}e^{-x\left(\frac{1-\beta t}{\beta}\right)}}{\Gamma(\alpha)\beta^\alpha} dx$$

setting $y = \frac{x(1-\beta t)}{\beta} \Rightarrow x = \frac{\beta y}{(1-\beta t)}$ and $dx = \frac{\beta}{(1-\beta t)}$

$$= \int_0^\infty \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)\beta^\alpha}\left(\frac{\beta}{1-\beta t}\right)^\alpha\left(\frac{1-\beta t}{\beta}\right)\left(\frac{\beta}{1-\beta t}\right)dy = \int_0^\infty \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\left(\frac{1}{1-\beta t}\right)^\alpha dy$$

$$= (1-\beta t)^{-\alpha}\int_0^\infty \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)} dy = (1-\beta t)^{-\alpha}$$

The Digamma function is defined as the derivative of the gamma function divided by the gamma function, mathematically written,

$$\text{Digamma function } \Psi(x) = \frac{\Gamma(x)'}{\Gamma(x)}$$

The likelihood function is found by finding the joint distribution of the product of the x_i's, mathematically written as,

$$L(\underline{x},\alpha,\beta) = \prod_{i=1}^n \frac{x_i^{\alpha-1}e^{-x_i/\beta}}{\Gamma(\alpha)\beta^\alpha} dx_i = \frac{\left(\prod_{i=1}^n x_i\right)^{\alpha-1} e^{\frac{\sum_{i=1}^n x_i}{\beta}}}{(\Gamma(\alpha))^n \beta^{n\alpha}} d\underline{x}$$

taking the log of this density yields the log-likelihood function,

$$\ln(L) = -n\ln\Gamma(\alpha) - n\alpha\ln(\beta) + (\alpha-1)\ln\left(\prod_{i=1}^n x_i\right) - \frac{\sum_{i=1}^n x_i}{\beta}$$

The maximum likelihood estimate (MLE) is obtained by finding the values of the parameters that maximize the above two equations. This is done by setting the partial

derivative of either of the above equations with respect to each parameter equal to zero and solving.

For the scale parameter β, using the log-likelihood function

$$\frac{\partial \ln(L)}{\partial(2\alpha)} = -\frac{n\alpha}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\alpha\beta \Rightarrow \hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}$$

and for the shape parameter α,

$$\frac{\partial \ln(L)}{\partial(\lambda/2)} = -\frac{n\Gamma(\alpha)'}{\Gamma(\alpha)} - n\ln\beta + \ln\left(\prod_{i=1}^{n} x_i\right) = 0, \text{ substituting } \beta = \frac{\bar{x}}{\alpha}$$

$$= -\frac{n\Gamma(\alpha)'}{\Gamma(\alpha)} - n\ln\left(\frac{\bar{x}}{\alpha}\right) + \ln\left(\prod_{i=1}^{n} x_i\right) = 0$$

$$= -\frac{n\Gamma(\alpha)'}{\Gamma(\alpha)} - n\ln\bar{x} + n\ln\alpha + \ln\left(\prod_{i=1}^{n} x_i\right) = 0$$

multiplying by $1/n$,

$$= -\frac{\Gamma(\alpha)'}{\Gamma(\alpha)} - \ln\bar{x} + \ln\alpha + \frac{\ln\left(\prod_{i=1}^{n} x_i\right)}{n} = 0$$

$$\Rightarrow \ln(\hat{\alpha}) - \Psi(\hat{\alpha}) = \ln\bar{x} - \sum_{i=1}^{n} \ln(x_i)/n$$

Newton-Raphson Method

The Newton-Raphson or Newtons method is suggested by Choi and Wette [1] and is derived as follows,

Suppose $f \in A^2[a,b]$, let $x_n \in [a,b]$ be an approximation to c such that $f(c) = 0$. From the defined space the first degree Taylor expansion, $P_2(x)$, expanded about x_n is,

$$P_2(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(x_n)}{2}(x - x_n)^2$$

expanding about c, given $f(c) = 0$, yields,

$$0 = f(x_n) + f'(x_n)(c - x_n) + \frac{f''(x_n)}{2}(c - x_n)^2$$

Newton's method assumes that $|c - x_n|$ is small and therefore $|c - x_n|^2$ is even smaller yielding,

$$0 \approx f(x_n) + f'(x_n)(c - x_n)$$

solving for c gives,

$c \approx x_n - \frac{f(x_n)}{f'(x_n)}$, this is the basis for the Newton-Raphson method, formally defined as,

$$c_n = c_{n-1} - \frac{f(c_{n-1})}{f'(c_{n-1})}$$

Newton's famous observation is that the second approximation to the zero of many functions is better than the preceding one, Berkey [2].

Convergence of Newton-Raphson Method

The convergence of a sequence is defined as follows,

Suppose $\{a_n\}_{n=0}^{\infty}$ is a sequence that converges to a, and $a_n \neq a$ for all n. then if constants exist such that,

$$\lim \frac{|a_{n=1} - a|}{|a_n - a|^{\alpha}} = \lambda, \text{ as } n \to \infty$$

then $\{a_n\}_{n=0}^{\infty}$ converges to a of the order α, with asymptotic error constant λ, Burden and Faires [3].

If $\alpha = 1$, the sequence is linearly convergent.

If $\alpha = 2$, the sequence is quadratically convergent.

To show that Newton's method is quadratically convergent the following theorem is given in Burden and Faires [3], p. 81.

Theorem 1: Let p be a solution of the equation $x = g(x)$. Suppose that $g'(p) = 0$ and g'' is continuous and strictly bounded by M on an open interval I containing p. Then there exists $\delta > 0$ such that, for $p_0 \in [p-\delta, p+\delta]$, the sequence defined by $p_n = g(p_{n-1})$, when $n \geq 1$, converges at least quadratically to p. Moreover, for sufficiently large values of n, $|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$.

To show that the Newton method is quadratically convergent is achieved by constructing a solution to the root-finding problem $f(x) = 0$, by subtracting a multiple of $f(x)$ from x, Burden and Faires [3]. The method is mathematically of the form,

$p_n = g(p_{n-1})$, where g is of the form $g(x) = x - \varphi(x)f(x)$

and $\varphi(x)$ is a differentiable function.

By definition g is quadratically convergent if $g'(p) = 0$. The derivative of the function $g(x)$ is,

$$g'(x) = 1 - \varphi'(x)f(x) - \varphi(x)f'(x)$$

using the above theorem, letting p be a solution of the equation $x = g(x)$, gives,

$$g'(p) = 1 - f'(p)\varphi(p), \text{ where } g'(p) = 0 \text{ if and only if } \varphi(p) = \frac{1}{f'(p)}$$

Therefore choosing $\varphi(x) = \dfrac{1}{f'(x)}$ produces the natural procedure that gives quadratic convergence as,

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \text{ or Newtons method, Burden and Faires [3].}$$

Application of Newtons Method for Finding the Maximum Likelihood Estimate of the Shape Parameter, $\hat{\alpha}$, of the Gamma Distribution

The equation for the MLE estimate, $\hat{\alpha}$, is given above as,

$$\ln(\alpha) - \Psi(\alpha) = \ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n \text{ or } \ln(\alpha) - \Psi(\alpha) - \left(\ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n \right) = 0$$

letting $Z = \left(\ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n \right)$, the equation is of the form,

$$f(\alpha) = \ln(\alpha) - \Psi(\alpha) - Z, \text{ and } f'(\alpha) = 1/\alpha - \Psi'(\alpha)$$

Therefore substituting into the formula for the Newton-Raphson iterative method,

$c_n = c_{n-1} - \dfrac{f(c_{n-1})}{f'(c_{n-1})}$, gives,

$$\hat{\alpha}_i = \hat{\alpha}_{i-1} - \frac{\ln \hat{\alpha}_{i-1} - \psi(\hat{\alpha}_{i-1}) - Z}{1/\hat{\alpha}_{i-1} - \psi'(\hat{\alpha}_{i-1})}. \quad (1)$$

The power series expansions for the derivative of the Digamma function, or Trigamma function, is $\Psi'(\alpha) = \sum_{i=0}^{\infty} (i+\alpha)^{-2}$, Choi and Wette [1] see Chapter 2, Section 1.

To ensure convergence, the denominator of (1) must be defined, i.e. $\alpha\Psi(\alpha) \neq 1$. Choi and Wette [1] note that $1/(x+\alpha)^2$ is a positive decreasing function of x, for $x \geq 0$. From this result the following inequality can be constructed,

$$\Psi'(\alpha)=\sum_{i=0}^{\infty}(i+\alpha)^{-2} > \int_{0}^{\infty}(x+\alpha)^{-2}dx=-(x+\alpha)^{-1}\Big|_{0}^{\infty}=1/\alpha$$

$\Rightarrow \Psi'(\alpha) > 1/\alpha$ or $\alpha\Psi(\alpha) > 1$, Choi and Wette [1].

Also the numerator and denominator of the last terms of (1) are monotone functions of α, therefore convergence is assured, Choi and Wette [1].

In Chapter, 2 Section 1, due to the quadratic convergence of this procedure, iteration was stopped after the absolute difference between $\hat{\alpha}_i$ and $\hat{\alpha}_{i-1}$ was less than 1×10^{-7}.

Bias of the Maximum Likelihood Estimates

Choi and Wette [1] performed a numerical study of the bias of the ML estimates by generating random samples from a Gamma distribution for $\alpha=1$, 2, 3, 5, and 7, $\beta=1$, of the sizes $n=40$, 120 and 200. For each case 100 independent repetitions were performed, with the tabulated results,

		$\hat{\alpha}$		$\hat{\beta}$	
α	n	mean	variance	Mean	variance
1	40	1.070	0.062	1.080	0.113
	120	1.030	0.010	1.040	0.022
	200	1.020	0.007	1.010	0.012
2	40	2.100	0.186	1.030	0.052
	120	2.040	0.073	1.020	0.019
	200	2.030	0.037	1.010	0.011
3	40	3.280	0.473	1.080	0.069
	120	3.170	0.155	1.060	0.019
	200	3.040	0.100	1.010	0.013
5	40	5.340	1.615	1.080	0.072
	120	5.190	0.492	1.040	0.023
	200	5.050	0.359	1.010	0.015
7	40	7.780	3.804	1.100	0.080
	120	7.200	0.933	1.030	0.021
	200	7.050	0.621	1.010	0.013

Table 3. 1 Means and Variances of $\hat{\alpha}$ and $\hat{\beta}$

From Table 3.1, we see that as the sample sizes increase the MLE estimates approach α, which is expected since the estimates are consistent. However, as Choi and Wette [1] point out, the sampling average of $\hat{\alpha}$ is greater than α in all examples. This gives strong evidence that the MLE estimates are positively bias, or $E(\hat{\alpha}) \geq \alpha$, Choi and Wette [1]. An adjustment for the bias was given in a paper by Johnson and Kotz [4] as

$$\hat{\alpha} = \hat{\alpha}(1-3/n)+2/3n \text{ for } n \geq 4 \text{and } \hat{\alpha} \geq 1 \text{ Schneider [5]}$$

Section 2: Upper Confidence Limit for the Mean of the Gamma Distribution

This section describes in more detail the material presented in Chapter 2, Section 2. When working with data from a contaminated site the mean of the distribution under

consideration is the most important characteristic to the experimenter. As Grice and Bain [6] point out standard estimates of the mean of a Gamma distribution based on a t distribution are often robust and useful. However, estimates based on this distribution become less appropriate for moderate values of α and small sample sizes, Grice and Bain [6]. In addition as the shape parameter becomes large the Gamma distribution approaches a normal distribution, Grice and Bain [6]. When analyzing contaminant concentration data with the Gamma distribution small sample sizes and moderate values of α are common and the reason the t distribution is not used. Grice and Bain [6] point out another good reason for not using the t distribution is that under a Gamma model the t statistic is based on moment estimates rather than the more efficient sufficient statistics, which make use of the data in the most proficient way. The upper confidence limit for the mean of a Gamma distribution is based on the uniformly most powerful test for testing the mean of a specified model say, μ_0 against a alternative composite hypothesis. As a result definitions are provided that lead to a more detailed look at the results obtained by Grice and Bain [6].

Definitions and Properties

Assuming the data $\{x_1, x_2, ..., x_n\}$, follows a Gamma distribution, $\Gamma(\alpha, \beta)$, then $Z = 2n\bar{x}/\beta$ follows a $\chi^2_{2n\alpha}$ distribution. This follows from,

$Y = \sum_{i=1}^{n} x_i$ has the moment generating function, $M_y(t) = E(e^{t\sum_{i=1}^{n} x_i}) = (1-\beta t)^{-n\alpha}$

then,

$$M_z(t) = E(e^{tz}) = E(e^{2t/\beta\sum_{i=1}^{n} x_i}) = (1-\beta 2t/\beta)^{-n\alpha} = (1-2t)^{-(2n\alpha)/2}$$

which is the moment generating function of a Chi-Square distribution with $2n\alpha$ degrees of freedom.

The transformation $K = \overline{X}/\alpha\beta$ follows a $\Gamma(n\alpha, 1/n\alpha)$ distribution. This is shown by deriving the moment generating function,

$$M_k(t) = E(e^{tk}) = E(e^{\frac{t}{n\alpha\beta}\sum_{i=1}^{n} x_i}) = \left(1 - \frac{1}{n\alpha}t\right)^{-n\alpha}.$$

Critical Region

A critical region is the determined subset of the sample space in which the assumed statistical hypothesis H_o is rejected.

Ex. Let the critical region be defined as, $C = \{(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 \geq 1\}$, the test defined by considering three random variables, X_1, X_2, X_3, and the observed values, x_1, x_2, x_3, then if $x_1^2 + x_2^2 + x_3^2 < 1$ accept H_o, or if $x_1^2 + x_2^2 + x_3^2 \geq 1$ reject H_o.

The notation $P[(x_1, x_2, ..., x_n) \in C, H_o]$ and $P[(x_1, x_2, ..., x_n) \in C, H_a]$ is the probability that the observed values are in the critical region, $P[(x_1, x_2, ..., x_n) \in C]$, under H_o and H_a respectively, Hogg and Craig [7].

The size of the test, or significance level is, $\alpha = P[(x_1, x_2, ..., x_n) \in C, H_o]$ and the power function is $P(x, \theta) = P[(x_1, x_2, ..., x_n) \in C]$.

Best Critical Region

Definition: Let $X \sim f(x,\theta)$, and $x_1, x_2, ..., x_n$ denote a random sample

Consider the test, $H_o : \theta = \theta'$ vs. $H_a : \theta = \theta''$, then the parameter space is $\Omega = \{\theta : \theta = \theta', \theta''\}$, and The Best Critical Region, C, of size α, in testing the simple hypothesis, is for every subset A of the sample space for which $P[(x_1, x_2, ..., x_n) \in A, H_o] = \alpha$ the following is true,

a.) $P[(x_1, x_2, ..., x_n) \in C, H_o] = \alpha$

b.) $P[(x_1, x_2, ..., x_n) \in C, H_a] \geq P[(x_1, x_2, ..., x_n) \in A, H_a]$, Hogg and Craig [22]

The best critical region occurs when the ratio, $f(x;\theta = \theta') / f(x;\theta = \theta'')$, has the smallest values of every subset A, of the sample space, satisfying the above conditions. A good example of this is provided in Hogg and Craig [7]. This leads to the following theorem by Neyman-Pearson for a systematic way of determining the best critical region.

Neyman-Pearson Theorem: Let $X \sim f(x,\theta)$, and $x_1, x_2, ..., x_n$ denote a random sample for a fixed n. Then the joint pdf of the random variables is $L(\underline{x},\theta) = f(\underline{x},\theta) = f(x_1,\theta), f(x_2,\theta), ..., f(x_n,\theta)$. Let θ' and θ'' be distinct fixed values of θ, so that the parameter space is, $\Omega = \{\theta : \theta = \theta', \theta''\}$, and let k be a positive number. Let C be a subset of the sample space under the following conditions,

a.) $\frac{L(\underline{x},\theta')}{L(\underline{x},\theta'')} \leq k$, for each point $x_i \in C$,

b.) $\frac{L(\underline{x},\theta')}{L(\underline{x},\theta'')} \geq k$, for each point $x_i \in C^c$

c.) $P[(x_1, x_2, ..., x_n) \in C, H_o] = \alpha$

Then C is a best critical region of size α for testing the simple hypothesis $H_o : \theta = \theta'$ vs. $H_a : \theta = \theta''$.

Examples using this theorem are found in most mathematical statistics textbooks, such as Hogg and Craig [7], p.401.

Uniformly Most Powerful Test

A uniformly most powerful critical region is just an extension of the definition and theorem previously stated to a test with a composite alternative hypothesis, for example $H_o : \theta = \theta'$ vs. $H_a : \theta > \theta'$. A test defined by this critical region is a uniformly most powerful test. The parameter space is now defined as, $\Omega = \{\theta : \theta > \theta'\}$. This is formally given in the following definition,

> Definition: The critical region C is a uniformly most powerful critical region of size α for testing H_o against a composite alternative hypothesis if the set C is a best critical region of size α for testing H_o against each simple alternative hypothesis in H_a. A test defined defined by this critical region C is called a uniformly most powerful test, with significance level α, for testing H_o against a composite alternative hypothesis H_a. Hogg and Craig [7]

An example of a (UMP) Test can be found in Hogg and Craig [7].

Upper Confidence Limit for the Mean of a Gamma Distribution Based on the Uniformly Most Powerful Test for $H_o : \alpha\beta \geq \alpha\beta_o$ vs. $H_a : \alpha\beta < \alpha\beta_o$.

The uniformly most powerful test of $H_o : \alpha\beta \geq \alpha\beta_o$ vs. $H_a : \alpha\beta < \alpha\beta_o$, when α is known is easily derived using the above stated definitions as follows;

(Note**this derivation is in Mood et al [8], or any of many mathematical texts)

Since α is known a simple hypothesis to test is $H_o : \alpha\beta = \alpha\beta'$ vs. $H_a : \alpha\beta = \alpha\beta''$. The best critical Region, C, letting $\beta' > \beta''$ and using the Neyman-Pearson Theorem is,

$$\frac{L(\underline{x},\theta')}{L(\underline{x},\theta'')} = \left(\frac{\beta''}{\beta'}\right)^{n\alpha} e^{\sum_{i=1}^{n} x_i\left(\frac{\beta'-\beta''}{\beta'\beta''}\right)} \leq k$$

$$\Rightarrow n\alpha \ln\left(\frac{\beta''}{\beta'}\right) + \sum_{i=1}^{n} x_i\left(\frac{\beta'-\beta''}{\beta'\beta''}\right) \leq \ln k$$

$$\Rightarrow \sum_{i=1}^{n} x_i \leq \frac{\beta'\beta''}{(\beta' - \beta'')}\left(\ln k - n\alpha \ln\left(\frac{\beta''}{\beta'}\right)\right) = c$$

$\therefore$ The best critical region is $C = \{(x_1, x_2, \ldots x_n) = \sum_{i=1}^{n} x_i < c\}$

In testing the hypothesis $H_o : \alpha\beta \geq \alpha\beta'$ vs. the composite alternative $H_a : \alpha\beta \leq \alpha\beta'$, let β''' be any value greater than β', than the above argument holds true.

$\therefore$ The uniformly most powerful critical region is $C = \{(x_1, x_2, \ldots x_n) = \sum_{i=1}^{n} x_i < c\}$

Since $2n\bar{x}/\beta$ follows a Chi-Square distribution a test based on the critical region $C = \{(x_1, x_2, \ldots x_n) = 2n\bar{x}/\beta < 2nc/\beta = c_1\}$ would be a uniformly most powerful test.

The size or significance level is,

$$\rho = P[(x_1 x_2, \ldots, x_n) \in C, H_o] = P(\bar{x}/\alpha\beta_o < \chi^2_{\rho(2n\alpha)}/2n\alpha) = \rho$$

and power function would follow as,

$$P(\theta) = P[(x_1, x_2, \ldots, x_n) \in C]$$

The symbol ρ is used in place of α, when talking of the significance level or confidence intervals to lessen any confusion with the shape parameter α of the Gamma distribution.

The probability statement for the size or significance level of the test can be rewritten as,

$$P(\bar{x}/\alpha\beta_o > \chi^2_{\rho(2n\alpha)}/2n\alpha) = 1 - \rho$$

$$= P(\alpha\beta_o < 2n\bar{x}\alpha/\chi^2_{\rho(2n\alpha)}) = 1 - \rho$$

$\therefore$ A $100(1-\rho)\%$ upper confidence limit is $2n\bar{x}\alpha/\chi^2_{\rho(2n\alpha)}$, and power function,

$$P(\theta) = P(\bar{x}/\alpha\beta_o < \chi^2_{\theta(2n\alpha)}/2n\alpha)$$

However, a problem exists in that the above results are based upon a known shape parameter, α, of the Gamma distribution. Grice and Bain [6] performed a numerical analysis of the power function for the test, $H_o : \alpha\beta \geq \alpha\beta_o$ vs. $H_a : \alpha\beta \leq \alpha\beta_o$ with α replaced by its MLE estimate $\hat{\alpha}$, discussed in Section 1 or Chapter 2, Section 1. The power function from Grice and Bain [6] is,

$$P_1(\alpha,\theta) = P(\bar{x}/\alpha\beta < \chi^2_{\theta(2n\hat{\alpha})}/2n\hat{\alpha})$$

where the probability statement is only dependent upon the unknown shape parameter and stochastically independent of $\bar{x}$ Grice and Bain [6].

Monte Carlo Study Estimated Values of $P_1(\alpha,\theta) = P(\bar{x}/\alpha\beta < \chi^2_{\theta(2n\hat{\alpha})}/2n\hat{\alpha})$

The Monte-Carlo simulation performed by Grice and Bain [6] for these values consisted of generating random samples of Gamma deviates with $\alpha\beta = 10$ at various combinations of n, α and $\beta = 10/\alpha$. The choice of $\alpha\beta = 10$ and β are immaterial in the computation of $P_1(\alpha,\theta)$ since, as shown in the properties section, $\bar{X}/\alpha\beta \sim \Gamma(n\alpha, 1/n\alpha)$, Grice and Bain [6]. From this and the fact that $\bar{x}$ and $\hat{\alpha}$ are independent, Grice and Bain [6] write the power function as,

$$P_1(\alpha,\theta) = \int_0^\infty P[\bar{x}/\alpha\beta < \chi^2_{\theta(2n\hat{\alpha})}/2n\hat{\alpha} | \hat{\alpha}] f(\hat{\alpha}) d\hat{\alpha}$$

letting $a = \hat{\alpha}$ gives,

$$= \int_0^\infty P[\bar{x}/\alpha\beta < \chi^2_{\theta(2na)}/2na] f(a) da$$

using the fact $\overline{X}/\alpha\beta \sim \Gamma(n\alpha,1/n\alpha)$, letting

$P[\overline{x}/\alpha\beta < \chi^2_{\theta(2na)}/2na] = G(\gamma = \chi^2_{\theta(2na)}/2na, n\alpha,1/n\alpha)$ where G is the cumulative

Gamma distribution, $G(\gamma,n\alpha,1/n\alpha) = \int_0^{\gamma} \Gamma(n\alpha,1/n\alpha)$, $\gamma = \left(\chi^2_{\theta(2na)}\right)^{-1}/2na$ gives,

$$= \int_0^{\infty} G(\gamma,n\alpha,1/n\alpha) f(a) da$$

The above statement is the expected value of G of $f(\hat{\alpha})$, $\hat{\alpha} = a$, therefore,

$$= E_{\hat{\alpha}}[G(\gamma,n\alpha,1/n\alpha)]$$

Grice and Bain [6] then generated between 2,000 to 4,000 samples for each combination of α and θ, computing $G(\chi^2_{\theta(2n\hat{\alpha})}/2n\hat{\alpha}, n\alpha,1/n\alpha)$ for each sample. These values were then averaged to obtain $P_1(\alpha,\theta)$.

Section 3: Confidence Interval for the Gamma Shape Parameter α

A more detailed explanation of the results obtained by Linhart [9], presented in Chapter 2 Section 3, for constructing a confidence interval for the shape parameter of the Gamma distribution.

Definitions and Properties

The Pearson Type III model is just the common Gamma model with shape parameter $\lambda/2$ and scale parameter 2α, and is used merely because it is the model chosen by both Linhart [9] and Bartlett [10]. The probability density function is,

$$f(x,\lambda/2,2\alpha) = \frac{x^{(\lambda/2)-1} e^{-x/2\alpha}}{\Gamma(\lambda/2)(2\alpha)^{\lambda/2}} dx,\ 0 \le x \le \infty,\ \lambda/2 \ge 0, 2\alpha \ge 0$$

with $\mu = \lambda\alpha$ and $\sigma = 2\alpha^2\lambda$

The moment generating function is obtained in the usual way by finding the expected value of e^{tx},

$$M_x(t) = E(e^{tx}) = \int_0^{\infty} e^{tx} \frac{x^{(\lambda/2)-1} e^{-x/2\alpha}}{\Gamma(\lambda/2)(2\alpha)^{\lambda/2}} dx = (1-2\alpha t)^{-\lambda/2}$$

The coefficient of variation for any probability density is the standard deviation divided by the mean. So for the Pearson Type III model the coefficient of variation is,

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{2\alpha^2\lambda}}{\alpha\lambda} = \frac{\sqrt{2\alpha}}{\lambda} = \sqrt{2}\lambda^{-1/2} = \left(\frac{2}{\lambda}\right)^{1/2}$$

The Digamma function is defined as the derivative of the gamma function divided by the gamma function, mathematically,

$$\text{Digamma function } \Psi(x) = \frac{\Gamma(x)'}{\Gamma(x)}$$

The Likelihood function is found by finding the joint distribution of the product of the x_i's, mathematically written as,

$$L(\underline{x}, \lambda/2, 2\alpha) = \prod_{i=1}^{n} \frac{x_i^{(\lambda/2)-1} e^{-x_i/2\alpha}}{\Gamma(\lambda/2)(2\alpha)^{\lambda/2}} dx_i = \frac{\left(\prod_{i=1}^{n} x_i\right)^{(\lambda/2)-1} e^{\frac{\sum_{i=1}^{n} x_i}{2\alpha}}}{(\Gamma(\lambda/n))^n (2\alpha)^{\frac{n\lambda}{2}}} d\underline{x}$$

taking the log of this density yields the log-likelihood function,

$$\ln(L) = -n\ln\Gamma(\lambda/2) - \frac{n\lambda}{2}\ln(2\alpha) + (\frac{\lambda}{2} - 1)\ln\left(\prod_{i=1}^{n} x_i\right) - \frac{\sum_{i=1}^{n} x_i}{2\alpha}$$

The maximum likelihood estimate (MLE) is obtained by finding the values of the parameters that maximize the above two equations. This is done by setting the partial

derivative of either of the above equations with respect to each parameter equal to zero and solving.

For the scale parameter 2α, using the log-likelihood function,

$$\frac{\partial \ln(L)}{\partial(2\alpha)} = -\frac{n\lambda}{2\alpha} + \frac{\sum_{i=1}^{n} x_i}{2\alpha^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\lambda\alpha \Rightarrow \alpha = \frac{\bar{x}}{\lambda}$$

and for the shape parameter $\lambda/2$,

$$\frac{\partial \ln(L)}{\partial(\lambda/2)} = -\frac{n\Gamma(\lambda/2)'}{\Gamma(\lambda/2)} - \frac{n\ln(2\alpha)}{2} + \frac{\ln\left(\prod_{i=1}^{n} x_i\right)}{2} = 0$$

$$= \frac{-2n\Gamma(\lambda/2)'}{\Gamma(\lambda/2)} - n\left(\ln 2 - \ln\lambda + \ln\bar{x}\right) + \ln\left(\prod_{i=1}^{n} x_i\right) = 0$$

multiplying by $1/n$,

$$= \frac{-2\Gamma(\lambda/2)'}{\Gamma(\lambda/2)} - \ln 2 + \ln\lambda - \ln\bar{x} + \frac{\ln\left(\prod_{i=1}^{n} x_i\right)}{n} = 0$$

$$= \frac{-2\Gamma(\lambda/2)'}{\Gamma(\lambda/2)} + \ln(\lambda/2) - \ln\bar{x} + \sum_{i=1}^{n} \ln(x_i)/n = 0$$

$$\Rightarrow \ln(\lambda/2) - \frac{2n\Gamma(\lambda/2)'}{\Gamma(\lambda/2)} = \ln\bar{x} - \sum_{i=1}^{n} \ln(x_i)/n$$

Notice the left side of the equation contains the Digamma function, and since no closed form solution exists its value must be obtained using numerical procedures. The numerical approximations for both $\Gamma(\lambda/2)'$ and $\Gamma(\lambda/2)$ have been analyzed and tabulated by Pairman [11].

In order to construct a confidence interval Linhart [9] uses Bartlett's [17] approximation defined as,

$$m = \ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n \approx \{[1+(1+1/n)/3\lambda]/n\lambda\}\chi^2_{n-1}$$

which yields the statistic

$$n\lambda m/[1+(1+n^{-1})/3\lambda] \approx \chi^2_{n-1}$$

Construction of the Confidence Interval

A 100(1-α)% confidence interval is obtained by rearrangement of the probability statement

$$P(\chi^2_{1-\alpha/2,n-1} \le \frac{n\lambda m}{[1+(1+n^{-1})/3\lambda]} \le \chi^2_{\alpha/2,n-1}) = 1-\alpha$$

$$\Rightarrow P\left(\frac{1}{\chi^2_{\alpha/2,n-1}} \le \frac{[1+(1+n^{-1})/3\lambda]}{n\lambda m} \le \frac{1}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{n\lambda m}{\chi^2_{\alpha/2,n-1}} \le [1+(1+n^{-1})/3\lambda] \le \frac{n\lambda m}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{n\lambda^2 m}{\chi^2_{\alpha/2,n-1}} \le \lambda + (1+n^{-1})/3 \le \frac{n\lambda^2 m}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{n\lambda^2 m}{\chi^2_{\alpha/2,n-1}} \le \lambda + \frac{(n+1)}{3n} \le \frac{n\lambda^2 m}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

The next step is to use the quadratic formula to solve both sides of the inequality inside the probability statement. The algebra is the same for both sides, therefore using just the left side we can rewrite the equality as,

$$\frac{n\lambda^2 m}{\chi^2_{\alpha/2,n-1}} - \lambda + \frac{(n+1)}{3n} \le 0$$

This equality is then solved with the quadratic equation, noting that $(\lambda - c)(\lambda + c) \leq 0 \Rightarrow (\lambda + c)$ is not a possible root since this would result in the statement $\lambda \leq -c$, which is invalid since $\lambda \geq 0$ by definition, resulting in the roots,

$$\left(\frac{1 \pm \sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1}}{2nm/\chi^2_{\alpha/2,n-1}} \right)$$

now, $m = \ln \bar{x} - \sum_{i=1}^{n} \ln(x_i)/n$ is positive since by definition a random variable x from a Gamma distribution is greater than or equal to 0. So the term $\sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1}$ has to be greater than 1. Therefore the only possible root is,

$$\left(\frac{1 + \sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1}}{2nm/\chi^2_{\alpha/2,n-1}} \right)$$

which leads to the inequality

$$\lambda \leq \left(\frac{1 + \sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1}}{2nm/\chi^2_{\alpha/2,n-1}} \right) \text{or } \lambda/2 \leq \left(\frac{1 + \sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1}}{4nm/\chi^2_{\alpha/2,n-1}} \right)$$

therefore the probability statement becomes

$$P\left(\frac{1 + (\sqrt{1 + 4(n+1)m/3}\chi^2_{1-\alpha/2,n-1})\chi^2_{1-\alpha/2,n-1}}{4nm} \leq \lambda/2 \leq \frac{1 + (\sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1})\chi^2_{\alpha/2,n-1}}{4nm} \right) = 1 - \alpha$$

Changing to the terms of the Gamma model used in Chapter 2, is done by substituting in the shape parameter $\alpha = \lambda/2$, giving the $(1-\alpha)\%$ confidence interval,

$$\left[\frac{1 + (\sqrt{1 + 4(n+1)m/3}\chi^2_{1-\alpha/2,n-1})\chi^2_{1-\alpha/2,n-1}}{4nm}, \frac{1 + (\sqrt{1 + 4(n+1)m/3}\chi^2_{\alpha/2,n-1})\chi^2_{\alpha/2,n-1}}{4nm} \right]$$

Section 4: Goodness OF FitTest for the Gamma Distribution

In this section the procedures and results of the paper written by Schneider [5] are discussed in more detail to provide more insight to the methods and results presented in Chapter 2, Section 4. The first part is merely a discussion of the procedures and methodology of the Monte Carlo study used and the methods and observations that lead to the smoothing function, $A(\rho)*\left(\frac{e^{B(\rho)\alpha^{-1/2}}}{n^{1/2}+C(p)}\right)$. This is followed by a discussion of the power of $D_n(\hat{\alpha},\hat{\beta})$, obtained by Schneider [5].

Definitions and Properties

Distribution functions and test statistics estimated from the data:

Gamma - $F_o(x,\hat{\alpha},\hat{\beta})$

EDF - $S_n(x)=\sum_{i=1}^{n}\varphi(x-x_i)/n$ where $\phi(z)=1$ for $z\geq 0$, 0 otherwise

Kolmogrov Smirnov type test statistic - $D_n(\hat{\alpha},\hat{\beta})=Sup_x\left|S_n(x)-F_o(x,\hat{\alpha},\hat{\beta})\right|$

where $D_n(\hat{\alpha},\hat{\beta})$ is used to denote the dependence of the statistic on the data size and estimated parameter values.

Hypothesis Being Tested - $Ho: F(x)=F_o(x,\hat{\alpha},\hat{\beta})$ *vs.* $Ha: F(x)\neq F_o(x,\hat{\alpha},\hat{\beta})$

for some x, where Ho is rejected for large values of D_n.

Monte Carlo Procedure Used in Calculating The Values of $D_n(\hat{\alpha},\hat{\beta})$

The range of the study performed by Schneider [5] was on data sizes of $n = 10,15,20,25,30$, $\alpha = 0.1,0.2,0.5,1,2,5,10,20,50$ at significance levels $\rho = 0.2,0.15,.0.1,0.05,0.01$. The statistic $D_n(\hat{\alpha},\hat{\beta})$ was estimated by generating a random sample of size n from a $\Gamma(\alpha,1)$ distribution to produce a single $D_n(\hat{\alpha},\hat{\beta})$. The parameter β was set to unity since the distribution of $D_n(\hat{\alpha},\hat{\beta})$ is independent of scale and location parameters, Schneider [5]. This procedure was then repeated 5000 times and the resulting $D_n(\hat{\alpha},\hat{\beta})$ were ordered producing the percentile points. The "raw" critical values are found in Schneider [5].

Smoothing Function for the Values of $D_n(\hat{\alpha},\hat{\beta})$

Examining these "raw" values Schneider [5] noticed for each case a fairly smooth contour over α and n at each significance level. As a point of reference Schneider [5] used the non-linear smoothing function for the exponential case presented in a paper by Stephens [12]. The function is of the form,

$$A^*(\rho)/(n^{1/2}+0.12+0.11n^{-1/2})$$

Schneider [5] started with functions of this form fitting them at each value of α using the MARQUARDT method in PROC NLIN of the Statistical Analysis System, (SAS), computing facility. The different values of the smoothing functions over n for each value of α were then plotted against values of α and visually reviewed, Schneider [5]. After reviewing some of the functional forms Schneider [5] decided to simultaneously fit the

"raw" values to both n and α. The results were, "after extensive experimentation, it was found that for a given significance level ρ, non-linear functions of the form

$$A(\rho) \quad * \left(\frac{e^{B(\rho)\alpha^{-1/2}}}{n^{1/2} + C(p)} \right)$$

fit the raw values reasonably well", Schneider [5]. It was found that this function worked well for $\alpha \geq 0.5$. Schneider [5] provides the estimated constants for the above smoothing function to be used for values of α not in the study. The table contains piecewise fits with break points between n=4(1)9 and n=10(5)30 which Schneider [5] found to be the most suitable in defining two distinct areas of each contour. In addition this function should also be used for extrapolation purposes for values of n outside the range of the study, Schneider [5]. In comparison to a paper by Durbin [13], who tabulated the exact null distribution of $D_n(\hat{\alpha}, \hat{\beta})$ for exponential data, the maximum deviation in results was 0.004 at $n = 4$ with an average difference less than 0.002, at the significance levels 0.05,0.10 and 0.20. In samples of size 10 and 30 the critical values at the same significance levels never differed by more than 0.001. Also noted for sample sizes of 50 and 100 use of the smoothing function provided similar results. Schneider [5]

Power of $D_n(\hat{\alpha}, \hat{\beta})$

In all of the studies below Schneider [5] generates 1,000 samples of size $n = 10$ and 500 samples of size $n = 20$ from a Gamma distribution using the procedure given in Chapter 2, Section 4. The first four sample moments for each of the 10,000 deviates generated are checked against the moments of the assumed underlying distribution to assure the samples were generated satisfactorily.

Study 1: Incorrectly specified shape parameters

In this study Schneider [5] investigates the power of $D_n(\hat{\alpha},\hat{\beta})$ by counting the number of times the null hypothesis is rejected for each generated sample at the significance level $\rho = 0.05$, for the values $\alpha = 0.1,0.2,0.5,1,2,5,10,20,50$.

For example Schneider [5] generates 1,000 samples of size $n = 10$ from a $\Gamma(0.1,1)$ distribution, calculates $D_n(\hat{\alpha},\hat{\beta})$ for each sample and tests the hypothesis $Ho: F(x) = F_o(x,\alpha,\beta)$ *vs.* $Ha: F(x) \neq F_o(x,\alpha,\beta)$ at each value $\alpha = 0.1,0.2,0.5,1,2,5,10,20,50$. The power of $D_n(\hat{\alpha},\hat{\beta})$ is then ascertained by counting the number of times Ho is rejected for each specified value $\alpha = 0.1,0.2,0.5,1,2,5,10,20,50$ and divided by 1,000. In this example when Ho is tested at $\alpha = 0.1$ the shape parameter is correctly specified and we expect the power to be equal to the size $\rho = 0.05$.

The following table provides the result of the Monte Carlo study of the power of $D_n(\hat{\alpha},\hat{\beta})$ for this study, Schneider [5].

Sample Size n	Underlying Distribution	α 0.1	0.2	0.5	1	2	5	10	20	50
10	Γ(0.1,1)	0.054	0.281	0.881	0.982	0.997	1.000	1.000	1.000	1.000
	Γ(0.2,1)	0.367	0.051	0.465	0.843	0.970	0.998	1.000	1.000	1.000
	Γ(0.5,1)	0.974	0.515	0.044	0.223	0.646	0.946	0.991	0.999	1.000
	Γ(1,1)	1.000	0.948	0.245	0.045	0.239	0.718	0.901	0.970	0.996
	Γ(2,1)	1.000	1.000	0.741	0.195	0.056	0.309	0.655	0.879	0.985
	Γ(5,1)	1.000	1.000	0.999	0.783	0.253	0.045	0.160	0.521	0.862
	Γ(10,1)	1.000	1.000	1.000	0.987	0.724	0.148	0.050	0.168	0.580
	Γ(20,1)	1.000	1.000	1.000	1.000	0.979	0.523	0.120	0.034	0.214
	Γ(50,1)	1.000	1.000	1.000	1.000	1.000	0.968	0.618	0.223	0.046
20	Γ(0.1,1)	0.054	0.556	0.992	1.000	1.000	1.000	1.000	1.000	1.000
	Γ(0.2,1)	0.662	0.048	0.754	0.994	1.000	1.000	1.000	1.000	1.000
	Γ(0.5,1)	1.000	0.836	0.050	0.486	0.948	1.000	1.000	1.000	1.000
	Γ(1,1)	1.000	1.000	0.514	0.066	0.414	0.934	0.990	1.000	1.000
	Γ(2,1)	1.000	1.000	0.980	0.394	0.050	0.502	0.908	0.994	1.000
	Γ(5,1)	1.000	1.000	1.000	0.994	0.560	0.042	0.274	0.780	0.992
	Γ(10,1)	1.000	1.000	1.000	1.000	0.984	0.294	0.034	0.248	0.858
	Γ(20,1)	1.000	1.000	1.000	1.000	1.000	0.896	0.288	0.050	0.412
	Γ(50,1)	1.000	1.000	1.000	1.000	1.000	1.000	0.966	0.392	0.058

Table 3. 2 Monte Carlo Study of the Power of $D_n(\hat{\alpha}, \hat{\beta})$

The diagonals of the table are the instances when the shape parameter is specified correctly, which means the power should be at the significance level, 0.05. For the samples of size 10 these values range from 0.034 to 0.056 with an average of 0.047 and from 0.034 to 0.066 with an average of 0.05 for the samples of size 20. Schneider [5] concludes that in both cases the significance levels are confirmed, with the differences being well within the limits of sampling error. As expected the power increases rapidly away from the correct shape parameter, Schneider [5].

Study 2: Both parameters unknown

In this study same procedure of study 1 is followed to calculate the powers of $D_n(\hat{\alpha},\hat{\beta})$ when the underlying distribution is Gamma and provides a check on the significance level $\rho = 0.05$. The results from Schneider [5] are tabulated as follows,

Underlying Distribution	Power of $D_n(\hat{\alpha},\hat{\beta})$ $n=10$	$n=20$
$\Gamma(0.1,1)$	0.069	0.066
$\Gamma(0.2,1)$	0.067	0.058
$\Gamma(0.5,1)$	0.048	0.056
$\Gamma(1,1)$	0.041	0.058
$\Gamma(2,1)$	0.049	0.050
$\Gamma(5,1)$	0.052	0.052
$\Gamma(10,1)$	0.046	0.042
$\Gamma(20,1)$	0.050	0.050
$\Gamma(50,1)$	0.056	0.064

Table 3. 3 Power of $D_n(\hat{\alpha},\hat{\beta})$ Under a Gamma Distribution

For the samples of size 10 the values range from 0.041 to 0.069 with an average of 0.053, and for sizes of 20 the values range from 0.042 to 0.066 with an average of 0.055. Schneider [5] concludes that the significance levels are reasonably well met.

Section 5: Generating Random Deviates

Given below is a step-by-step derivation of the result given by Whittaker [14], presented in Chapter 2, Section 5, for creating a random variable from a Gamma distribution, $f(x,\alpha,\beta)=\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}dx$.

Definitions

$M_x(t)=E(e^{tx})=(1-\beta t)^{-\alpha}$

Uniform Distribution: $f(u)=du$, $0<u<1$.

Beta Distribution: $f(x,p,q)=\beta(p,q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1}dx$, $0<x<1$

Result 1

The following is used for producing random deviates from both a Gamma and Beta distribution.

$X=-\beta\sum_{i=1}^{[\alpha]}\ln U_i \sim \Gamma([\alpha],\beta)$, $U_i \sim U(0,1)$ and, $[\alpha]$ is the integer part of α.

Proof:

Letting $\alpha=1$ we have the random variable $X=-\beta\ln U \Rightarrow U=e^{-x/\beta}$ and $du=-\frac{e^{-x/\beta}}{\beta}dx$. Since U is a uniform random variable defined above, using the change of variable technique the distribution of X is,

$f(x)=\frac{e^{-x/\beta}}{\beta}dx$, which is $\Gamma(1,\beta)$

letting $Y = \sum_{i=1}^{k} X_i$ where k is any integer from 1 to ∞, then the distribution of Y, using the moment generating function technique, is,

$$M_y(t) = E(e^{ty}) = E\left(e^{t\sum_{i=1}^{k} x_i}\right) = E(e^{tx_1})E(e^{tx_2})...E(e^{tx_k}) = (1-\beta t)^{-k}$$

which is the moment generating function for a Gamma distribution, $\Gamma(k,\beta)$.

$\therefore X = -\beta\sum_{i=1}^{[\alpha]} \ln U_i$ follows a Gamma distribution, $\Gamma([\alpha],\beta)$ //

Result 2

The following derivation is outlined in Whittaker [14] and can be used to produce random deviates from a Beta distribution and is a basis for producing Gamma deviates.

$Y = \frac{S_1}{S_1 + S_2}$ follows a $\beta(p,q)$, where $S_1 = U_1^{1/p}$, $S_2 = U_2^{1/q}$ and U_1, U_2 are independent $U(0,1)$ variables, if and only if $S_1 + S_2 \le 1$

Proof:

The joint distribution of U_1 and U_2 is $f(u_1,u_2) = f(u_1)f(u_2) = du_1 du_2$, $0 < u_1, u_2 < 1$. The joint distribution of S_1 and S_2, is obtained by determining the Jacobian as follows,

$$J = \begin{vmatrix} \frac{\partial u_1}{\partial s_1} & \frac{\partial u_1}{s_2} \\ \frac{\partial u_2}{\partial s_1} & \frac{\partial u_2}{\partial s_2} \end{vmatrix} = \begin{vmatrix} ps_1^{p-1} & 0 \\ 0 & qs_2^{q-1} \end{vmatrix} = ps_1^{p-1}qs_2^{q-1} \Rightarrow$$

$$f(s_1,s_2) = ps_1^{p-1}qs_2^{q-1}\, ds_1 ds_2,\ 0 < s_1, s_2 < 1$$

or can be obtained as follows,

$$ds_1 = \frac{u_1^{\frac{1-p}{p}}}{p} du_1 \Rightarrow du_1 = p u_1^{\frac{p-1}{p}} ds_1 = p s_1^{p-1} ds_1 \text{ and}$$

$$ds_2 = \frac{u_2^{\frac{1-q}{q}}}{q} du_2 \Rightarrow du_2 = q u_2^{\frac{q-1}{q}} ds_2 = q s_2^{q-1} ds_2$$

by substitution into $f(u_1, u_2) \Rightarrow f(s_1, s_2) = p s_1^{p-1} q s_2^{q-1} ds_1 ds_2,\ 0 < s_1, s_2 < 1$

now define the random variables $Y = \frac{S_1}{S_1 + S_2}$ and $Z = S_1 + S_2$ then,

$$Y = \frac{S_1}{S_1 + S_2} \Rightarrow (S_1 + S_2)Y = S_1 \Rightarrow S_1 = YZ \text{ and}$$

$$Z = S_1 + S_2 \Rightarrow S_2 = Z - YZ = Z(1 - Y)$$

the Jacobian is,

$$J = \begin{vmatrix} \frac{\partial s_1}{\partial y} & \frac{\partial s_1}{\partial z} \\ \frac{\partial s_2}{\partial y} & \frac{\partial s_2}{\partial z} \end{vmatrix} = \begin{vmatrix} z & y \\ -z & (1-y) \end{vmatrix} = z(1-y) + zy = z$$

therefore the joint distribution of Y and Z is,

$$f(y, z) = pq y^{p-1} (1-y)^{q-1} z^{p+q-1} dy dz$$

bounded by,

$$S_1 + S_2 < 2 \Rightarrow Z \le 2,\ Z \ge 0,\ Z \le \frac{1}{Y} \text{ and } Z \le \frac{1}{1-Y}$$

Result 3

Producing random Gamma deviates is outlined in Whittaker [14] as follows using the results from above. Again let U_i be an independent $U(0,1)$ random variable and define,

$$Y = \frac{S_1}{S_1 + S_2}, \text{ where } S_1 = U_1^{1/p},\ S_2 = U_2^{1/1-p},\ S_1 + S_2 \le 1 \text{ and } X = -Y \ln U_3$$

the joint distribution of X and Y is,

$$f(x,y) = \frac{\Gamma[p+(1-p)]}{\Gamma(p)\Gamma(1-p)} e^{-x} y^{p-1}(1-y)^{(1-p)-1} = \frac{e^{-x} y^{p-1}(1-y)^{-p}}{\Gamma(p)\Gamma(1-p)} dxdy$$

Defining the following transformations,

$$Z_1 = XY \Rightarrow X = \frac{Z_1}{Y} \text{ and } Z_2 = (1-Y)X \Rightarrow Y = 1 - \frac{YZ_2}{Z_1} = \frac{1}{(1+Z_2/Z_1)}$$

$$\Rightarrow Y = \frac{Z_1}{Z_1 + Z_2} \text{ and } X = Z_1 + Z_2$$

therefore the Jacobian is,

$$J = \begin{vmatrix} \frac{\partial x}{\partial z_1} & \frac{\partial x}{\partial z_2} \\ \frac{\partial y}{\partial z_1} & \frac{\partial y}{\partial z_2} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ \frac{z_2}{(z_1+z_2)^2} & \frac{-z_1}{(z_1+z_2)^2} \end{vmatrix} = \begin{vmatrix} \frac{-z_1}{(z_1+z_2)^2} - \frac{z_2}{(z_1+z_2)^2} \end{vmatrix} = (z_1+z_2)^{-1}$$

The joint distribution of Z_1 and Z_2 is derived as follows,

$$f(z_1,z_2) = \frac{1}{\Gamma(p)\Gamma(1-p)} e^{-(z_1+z_2)} \left(\frac{z_1}{z_1+z_2}\right)^{p-1} \left(1 - \frac{z_1}{z_1+z_2}\right)^{-p} (z_1+z_2)^{-1} dz_1 dz_2$$

$$= \frac{1}{\Gamma(p)\Gamma(1-p)} e^{-z_1} e^{-z_2} z_1^{p-1} \left(\frac{1}{z_1+z_2}\right)^{p} \left(\frac{1}{z_1+z_2}\right)^{-1} \left(\frac{z_2}{z_1+z_2}\right)^{-p} (z_1+z_2) dz_1 dz_2$$

$$= \frac{e^{-z_1} e^{-z_2} z_1^{p-1} z_2^{-p}}{\Gamma(p)\Gamma(1-p)} dz_1 dz_2 = \frac{e^{-z_1} z_1^{p-1}}{\Gamma(p)} dz_1 \cdot \frac{e^{-z_2} z_2^{-p}}{\Gamma(1-p)} dz_2$$

Since the joint density is the product of two Gamma densities the transformed variables are independent, therefore,

$$Z_1 = XY \Rightarrow Z_1 = -\left(\frac{S_1}{S_1 + S_2}\right)\ln U_3 \sim \Gamma(p,1) \text{ and}$$

$$Z_2 = (1-Y)X \sim \Gamma(1-p,1)$$

Note from earlier we have $X = -\beta\sum_{i=1}^{[\alpha]} \ln U_i \sim \Gamma([\alpha],\beta)$, and therefore multiplying Z_1 by β would follow a Gamma distribution, $\Gamma(p,\beta)$.

Finally to produce a random Gamma deviate with both scale and shape parameters define $p = \alpha + [\alpha]$ and use the transformation $Z = X + Z_1$, which follows a $\Gamma(\alpha,\beta)$ distribution.

$$\because M_z(t) = E(e^{tz}) = E(e^{t(x+z_1)}) = E(e^{tx})E(e^{tz_1}) = (1-\beta t)^{-p}(1-\beta t)^{-[\alpha]}$$

$$= (1-\beta t)^{-(p+[\alpha])} = (1-\beta t)^{-\alpha} \sim \Gamma(\alpha,\beta)$$

The subroutine for producing random variables for a Gamma distribution can be found in Chapter 4, subroutine Gammadeviates. This subroutine uses the function Ran1, Press et al [15], to produce random variables from a Uniform distribution.

Section 6: The Lognormal Distribution and the H-Statistic

This section shows detailed derivations of the MVUE estimate for the mean of the lognormal distribution given by Bradu and Mundlak [16], presented in Chapter 2, Section 6. Also discussed is Cubic Lagrangian interpolation on the H tables.

lognormal density,

Let $X = \{x_1, x_2, ..., x_n\}$ represent a random sample, then if $Y = \ln X$ follows a $N(\mu, \sigma^2)$ distribution $X = e^Y$ follows a $LN(\mu, \sigma^2)$ distribution with probability density function,

$$f_x = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(\ln x - \mu)^2} \frac{1}{x} dx, \quad 0 \le x \le \infty,$$

The moment generating function for Y is $M_y(t) = E(e^{ty}) = e^{\mu t + 1/2\sigma^2 t^2}$, therefore the moments of X can be derived as follows,

Since $E(X^k) = E(e^{ky}) = e^{\mu k + 1/2\sigma^2 k^2}$ we have

mean is $\mu_1 = E(X) = E(X^{k=1}) = e^{\mu + 1/2\sigma^2}$

Variance is $\sigma_1^2 = V(X) = E(X^{k=2}) - (E(X^{k=1}))^2 = e^{2\mu+\sigma^2}[e^{\sigma^2} - 1]$

The moment generating function of a Chi-Square distribution with m degrees of freedom, denoted χ_m^2, is,

$$M_x(t) = E(e^{tx}) = (1-2t)^{-m/2}$$

$$E(\chi_m^2) \text{ is } M_x'(t) = m(1-2t)^{-(m/2+1)}\Big|_{t=0} = m$$

$$E(\chi_m^2)^2 \; M_x''(t) = (m+2)m(1-2t)^{-(m/2+1)}\Big|_{t=0} = (m+2)m$$

Theorem 1. Let $f(x,\theta)$, $\gamma < \theta < \delta$, be a pdf. which represents a regular case of the exponential class. Then if $X_1, X_1, ..., X_n$, where n is a fixed positive integer, is a random sample from a distribution with pdf. $f(x,\theta)$, the statistic $Y_1 = \sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for θ and the family $\{g_1(y_1;\theta) : \gamma < \theta < \delta\}$ of probability density functions of Y_1 is

complete. That is, Y_1 is a complete sufficient statistic for θ.
(1) if we can see a function of the form Y_1, say $\varphi(Y_1)$, such that $E[\varphi(Y_1)] = \theta$, then the statistic $\varphi(Y_1)$ is unique and is the minimum variance unbiased estimate, (MVUE). Hogg and Craig [7].

Minimum Variance Unbiased Estimate for the Mean of A Lognormal Distribution

Bradu and Mundlak [16] derived the unbiased estimate with the minimum variance of all unbiased estimates, denoted MVUE, by expanding on results from a paper written by Finney [17]. The following is a detailed look at the results obtained by Bradu and Mundlak [16] using the 'Finney Solution'.

Let (z, s^2) be two independent random variables where $Z = \ln X \sim N(\xi, \upsilon\sigma^2)$ and $S^2 \sim \sigma^2\chi_n^2 / n$. The moments of S^2, denoted $E(\sigma^2\chi_n^2 / n)^k$, can be represented in the functional form as follows,

$$E(s^{2k}) = \frac{n(n+2)...(n+2k)}{n^k(n+2k)}\sigma^{2k},\ k = 0,1,.....\infty$$

This is seen in the following comparison of the first two terms using moment generating function of a Chi-Square distribution given above,

$k = 1$

$$E(S^{2k}) = E(S^2) = \frac{n(n+2)}{n(n+2)}\sigma^2 = \sigma^2$$

or

$$E(\sigma^2\chi_n^2 / n)^k = \frac{\sigma^2}{n}E(\chi_n^2) = \frac{\sigma^2}{n}n = \sigma^2$$

$k = 2$

$$E(S^{2k}) = E(S^2)^2 = \frac{n(n+2)(n+4)}{n^2(n+4)}(\sigma^2)^2 = \frac{(n+2)}{n}(\sigma^2)^2$$

or

$$E(\sigma^2\chi_n^2/n)^k = \frac{(\sigma^2)^2}{n^2}E(\chi_n^2)^2 = \frac{(\sigma^2)^2}{n^2}n(n+2) = \frac{(n+2)}{n}(\sigma^2)^2$$

Finney [17] introduced the following function,

$$g_n(t) = \sum_{k=0}^{\infty} w_k(t) = \sum_{k=0}^{\infty} \frac{n^k(n+2k)}{n(n+2)...(n+2k)}\left(\frac{n}{n+1}\right)^k \frac{1}{k!}t^k$$

which is combined with the previous equation and algebraically simplified,

$$E(g_n(As^2)) = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{n^k(n+2k)}{n(n+2)..(n+2k)}\left(\frac{n}{n+1}\right)^k A^k E(S^2)^k$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!}\frac{n^k(n+2k)}{n(n+2)..(n+2k)}\left(\frac{n}{n+1}\right)^k A^k \frac{n(n+2)...(n+2k)}{n^k(n+2k)}(\sigma^2)^k$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{n}{n+1}A\sigma^2\right)^k = e^{nA\sigma^2/(n+1)}$$

Therefore the function, $g_n(As^2)$ is an unbiased estimate for $e^{nA\sigma^2/(n+1)}$.

By definition the transformation $X = e^{BZ}$ must follow a lognormal distribution, since $ZB \sim N(B\xi, \upsilon\sigma^2B^2)$, and $E(Y) = E(e^{Bz}) = e^{B\xi + B^2\sigma^2\upsilon/2}$.

$\therefore$ e^{Bz} is an unbiased estimate of $e^{B\xi+B^2\sigma^2\upsilon/2}$.

Combining results we have,

$$E(e^{Bz}g_n(As^2)) = E(e^{Bz})E(g_n(As^2)) = e^{B\xi+B^2\sigma^2\upsilon/2}\, e^{nA\sigma^2/(n+1)} = e^{B\xi+(B^2\upsilon/2+\frac{n}{n+1}A)\sigma^2}$$

$\therefore$ $e^{Bz}g_n(As^2)$ is an unbiased estimate of $e^{B\xi+(B^2\upsilon/2+\frac{n}{n+1}A)\sigma^2}$

To achieve an unbiased estimate for a function of the form, $e^{\tau\xi+c\sigma^2}$, we must find a solution to the following,

$$B\xi+B^2\sigma^2\upsilon/2+\frac{n}{n+1}A\sigma^2=\tau\xi+c\sigma^2, \text{ letting } B=\tau \text{ gives,}$$

$$\tau\xi+\tau^2\sigma^2\upsilon/2+\frac{n}{n+1}A\sigma^2=\tau\xi+c\sigma^2\Rightarrow\tau^2\sigma^2\upsilon/2+\frac{n}{n+1}A\sigma^2=c\sigma^2$$

$$\Rightarrow\tau^2\upsilon/2+\frac{n}{n+1}A=c\Rightarrow A=\frac{n+1}{n}\left(c-\tau^2\upsilon/2\right)$$

$$\therefore E[e^{\tau}g_n(\frac{n+1}{n}(c-\tau^2\upsilon/2)s^2)]=e^{\tau\xi+(\tau^2\upsilon/2+c-\tau^2\upsilon/2)\sigma^2}=e^{\tau\xi+c\sigma^2}$$

so $e^{\tau}g_n(\frac{n+1}{n}(c-\tau^2\upsilon/2)s^2)$ is an unbiased estimate of $e^{\tau\xi+c\sigma^2}$

Using this result with the estimates,

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\sim\chi^2_{n-1} \text{ and } \bar{x}\sim N(\tau\mu,\tau^2\sigma^2/n)$$

$\Rightarrow$ $z=\bar{x}$, $n=n-1$, $\xi=\mu$ and $\upsilon=1/n$

which yields,

$$E[e^{\bar{x}}g_n\left(\frac{n}{n-1}(c-\tau^2/2n)s^2\right)]=E[e^{\bar{x}}g_n\left(\frac{2nc-\tau^2}{2(n-1)}s^2\right)]=e^{\tau\mu+\tau^2\sigma^2/2n+\left(\frac{n-1}{n}\left(\frac{2nc-\tau^2}{2(n-1)}\right)\right)\sigma^2}$$

$$=e^{\tau\mu+\tau^2\sigma^2/2n+(c-\tau^2/2n)\sigma^2}=e^{\tau\mu+\sigma^2c}$$

$\therefore$ $e^{\bar{x}}g_n\left(\frac{2nc-\tau^2}{2(n-1)}s^2\right)$ is an unbiased estimate of $e^{\tau\mu+\sigma^2c}$

This is the uniformly minimum variance unbiased estimator for the estimated value since $\bar{x}$ and s^2 are jointly sufficient and complete statistics. This is shown using Theorem 1, which can be extended to the case of several parameters, Hogg and Craig [7]. The proof

is based on $Z = \ln X \sim N(\theta_1, \theta_2)$, which implies $X = e^Z \sim LN(\theta_1, \theta_2)$. We have,

$$f(x,\theta_1,\theta_2) = (2\pi\theta_2)^{-1/2} e^{-\frac{1}{2\theta_2}(\ln x - \theta_1)^2} \frac{1}{x} dx$$

which can be written in the exponential form,

$$f(x,\theta_1,\theta_2) = e^{-\frac{1}{2\theta_2}(\ln x)^2 + \frac{\theta_1}{\theta_2}\ln x - \frac{\theta_1^2}{2\theta_2} - \ln[(2\pi\theta_2)^{1/2} - \ln x]} dx = e^{-\frac{1}{2\theta_2}(\ln x)^2 + \left(\frac{\theta_1}{\theta_2} - 1\right)\ln x - \frac{\theta_1^2}{2\theta_2} - \ln[(2\pi\theta_2)^{1/2}]} dx$$

so we have $K_1(x) = (\ln x)^2$ and $K_2(x) = \ln x$. Therefore, the statistics,

$$Y_1 = \sum_{i=1}^{n} (\ln x_i)^2 \text{ and } Y_2 = \sum_{i=1}^{n} \ln x_i$$

are joint complete and sufficient statistics for θ_1 and θ_2. Defining the following one-to-one transformations,

$$Z_1 = \frac{Y_2}{n} = \ln \bar{x} \text{ and } Z_2 = \frac{Y_1 - Y_2^2}{n-1} = \frac{\sum_{i=1}^{n} (\ln x_i - \ln \bar{x})^2}{n-1} = s^2$$

are also joint and complete sufficient statistics. Therefore, we have,

$$e^{\bar{z}} g_n\left(\frac{2nc - \tau^2}{2(n-1)} s^2\right) \text{ is the MVUE of } e^{\tau\mu + \sigma^2 c}$$

Useful Properties of the Lognormal Distribution Using the Above Results

Again, $Z = \ln X \sim N(\mu, \sigma^2)$ and $X = e^Z \sim LN(\mu, \sigma^2)$, $\bar{z}$ and s_z are based on the log-transformed data, $z_i = \ln(x_i)$, $i = 1,2,...,n$.

Mean $E(X) = \mu_1 = e^{\mu + 1/2\sigma^2}$, $\Rightarrow \tau = 1$ and $c = 1/2$

$$\Rightarrow E[e^{\bar{z}} g_n\left(\frac{n - \tau^2}{2(n-1)} s^2\right)] = e^{\mu + \sigma^2/2n + \left(\frac{n-1}{n}\left(\frac{n-\tau^2}{2(n-1)}\right)\right)\sigma^2} = e^{\mu + \sigma^2/2n + \left(\frac{n-\tau^2}{2n}\right)\sigma^2} = e^{\mu + 1/2\sigma^2}$$

$$\therefore \hat{\mu}_1 = e^{\bar{z}} g_n\left(\frac{n-\tau^2}{2(n-1)}s^2\right) \text{is the MVUE of } e^{\mu+1/2\sigma^2}$$

Median $X = e^{\mu} \Rightarrow \tau = 1$ and $c = 0$

$$\therefore e^{\bar{z}} g_n\left(\frac{-\tau^2}{2(n-1)}s^2\right) \text{ is the MVUE of } e^{\mu}$$

$E(X^k) = e^{\mu k+1/2\sigma^2 k^2} \Rightarrow \tau = k$ and $c = k^2/2$

$$\Rightarrow E[e^{k\bar{z}} g_n\left(\frac{nk^2-k^2}{2(n-1)}s^2\right)] = E[e^{k\bar{z}} g_n\left(\frac{k^2}{2}s^2\right)] = e^{k\mu+k^2\sigma^2/2n+\left(\frac{n-1}{n}\left(\frac{k^2}{2}\right)\right)\sigma^2}$$

$$= e^{k\mu+k^2\sigma^2/2n+\left(\frac{k^2}{2}-\frac{k^2}{2n}\right)\sigma^2} = e^{\mu k+1/2\sigma^2 k^2}$$

$E(X)^k = e^{\mu k+1/2\sigma^2 k} \Rightarrow \tau = k$ and $c = k/2$

$$\Rightarrow E[e^{k\bar{z}} g_n\left(\frac{k(n-k)}{2(n-1)}s^2\right)] = e^{k\mu+k^2\sigma^2/2n+\left(\frac{n-1}{n}\left(\frac{k(n-k)}{2(n-1)}\right)\right)\sigma^2}$$

$$= e^{k\mu+k^2\sigma^2/2n+\frac{\sigma^2 k}{2}-\frac{k^2\sigma^2}{2n}} = e^{\mu k+1/2\sigma^2 k}$$

Variance $V(X) = \sigma_1^2 = E(X^2) - (E(X))^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$, using the formulas we have,

$$E(X^2) = E[e^{2\bar{z}} g_n(2s^2)] \text{ and } (E(X))^2 = E[e^{2\bar{z}} g_n\left(\frac{(n-2)}{(n-1)}s^2\right)]$$

therefore the MVUE estimate for the variance is,

$$\hat{\sigma}_1^2 = e^{2\bar{z}} g_n(2s^2) - e^{2\bar{z}} g_n\left(\frac{(n-2)}{(n-1)}s^2\right) = e^{2\bar{z}}\left[g_n(2s^2) - g_n\left(\frac{(n-2)}{(n-1)}s^2\right)\right]$$

The variance of the estimate, $\hat{\mu}_1$, is also given by Bradu and Mundlak [16] as follows,

$$\hat{\sigma}^2_{\hat{\mu}_1} = e^{2\bar{y}\left[\left(g_n(s_y^2/2)\right)^2 - g_n\left((n-2)s_y^2/(n-1)\right)\right]}$$

Lagrangian Interpolation for Tables of the H-Statistic

As mentioned in Chapter 2, Section 6, in environmental applications the most commonly used upper bound for the mean of a lognormal distribution is based on the H statistic given by Land [18][19]. Tabulated values for the H-Statistic are found in Land [18][19] and Gilbert [20] and are based on s and n. When either one or both of these values are not on the table, Land [28] suggests Lagrangian Cubic Interpolation.

Lagrangian Cubic Interpolation is a technique used to fit an unknown function with a continuous polynomial through the known data points. Gerald and Wheatley [21] give the functional form as,

$$H_3(x) = \frac{(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)} f_1 + \frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)} f_2 + \frac{(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)} f_3 + \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_1)(x_4-x_2)(x_4-x_3)} f_4$$

The process is to omit one x_i value and write the numerator as, $(x-x_i)$, where x_i is a non-omitted value. The denominator is the product of the distance each of these points is from the omitted value, $(x_{i(omitted)}-x_i)$. The function itself is made up of four terms, each of which is a cubic in x, which sum to a cubic, Gerald and Wheatley [21]. This procedure is illustrated in the following example, where s_y is not a value present on the table,

Ex. $n=3$, $s_y=3.5$ and we want to interpolate for H. The corresponding table values are,

s_y	$n=3$
0.2	3.295
0.3	4.109
0.4	5.22
0.5	6.495

Table 3. 4 Partial Table of H-Values

$$H_3(.35) =$$

$$\frac{(3.5-.3)(3.5-.4)(3.5-.5)}{(.2-.3)(.2-.4)(.2-.5)}3.295+\frac{(3.5-.2)(3.5-.4)(3.5-.5)}{(.3-.2)(.3-.4)(.3-.5)}4.109.... = 5.032926$$

In cases where s_y and n are both not represented in the table the above formula will be adapted by performing cubic Lagrangian interpolation in the x direction followed by cubic Lagrangian interpolation in the y direction.

$$H_3(x,y)=\frac{(y-y_2)(y-y_3)(y-y_4)}{(y_1-y_2)(y_1-y_3)(y_1-y_4)}\Big[\ \frac{(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)}f_1+$$

$$\frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}f_2+\frac{(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)}f_3+$$

$$\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_1)(x_4-x_2)(x_4-x_3)}f_4\big|y_1\big]+....+\frac{(y-y_1)(y-y_2)(y-y_3)}{(y_4-y_1)(y_4-y_2)(y_4-y_3)}$$

$$\Big[\ \frac{(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)}f_1+\frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}f_2+$$

$$\frac{(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)}f_3+\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_1)(x_4-x_2)(x_4-x_3)}f_4\big|y_4\big]$$

An example for this is given in Gerald and Wheatley [21],

Ex. Cubic Lagrangian interpolation in the y direction and quadratic Lagrangian interpolation in the x direction.

	y			
x	0.2	0.3	0.4	0.5
1	0.64	1.003	1.359	1.703
1.5	0.99	1.524	2.045	5.549
2	1.568	2.384	3.177	3.943

Table 3. 5 Partial Table of H-Values

The Lagrangian equation for any x and y value would be

$$H(x,y)=\frac{(y-0.3)(y-0.4)(y-0.5)}{(0.2-0.3)(0.2-0.4)(0.2-0.5)}\left[\frac{(x-1.5)(x-2.0)}{(1.0-1.5)(1.0-2.0)}(0.64)+\right.$$

$$\left.\frac{(x-1.0)(x-2.0)}{(1.5-1.0)(1.5-2.0)}(0.99)+\frac{(x-1.0)(x-1.5)}{(2.0-1.0)(2.0-1.5)}(1.568)\right]+$$

$$\frac{(y-0.2)(y-0.4)(y-0.5)}{(0.3-0.2)(0.3-0.4)(0.3-0.5)}\left[\frac{(x-1.5)(x-2.0)}{(1.0-1.5)(1.0-2.0)}(1.003)\right.$$

$$\left.+\frac{(x-1.0)(x-2.0)}{(1.5-1.0)(1.5-2.0)}(1.534)+\frac{(x-1.0)(x-1.5)}{(2.0-1.0)(2.0-1.5)}(2.384)\right]+$$

$$\frac{(y-0.2)(y-0.3)(y-0.5)}{(0.4-0.2)(0.4-0.3)(0.4-0.5)}\left[\frac{(x-1.5)(x-2.0)}{(1.0-1.5)(1.0-2.0)}(1.359)\right.$$

$$\left.+\frac{(x-1.0)(x-2.0)}{(1.5-1.0)(1.5-2.0)}(2.045)+\frac{(x-1.0)(x-1.5)}{(2.0-1.0)(2.0-1.5)}(3.177)\right]+$$

$$\frac{(y-0.2)(y-0.3)(y-0.4)}{(0.5-0.2)(0.5-0.3)(0.5-0.4)}\left[\frac{(x-1.5)(x-2.0)}{(1.0-1.5)(1.0-2.0)}(1.703)\right.$$

$$\left.+\frac{(x-1.0)(x-2.0)}{(1.5-1.0)(1.5-2.0)}(2.549)+\frac{(x-1.0)(x-1.5)}{(2.0-1.0)(2.0-1.5)}(3.943)\right]+$$

Note in the results of Chapter 2, all 2 dimensional Interpolation uses cubic Lagrangian interpolation in both directions.

Section 7: Bootstrapping

This section provides more details and explanations of the bootstrapping methods presented in Chapter 2. The topics deal with the nonparametric method of bootstrapping since the results can be applied to parametric methods. Bootstrapping is just a name for the procedure of resampling from the original data either directly or from a fitted model. This section focuses on the methods for resampling from the original data set. The procedure involves in its simplest formulation generating replicate data sets based on the original data set, these replicated sets are generated by a computer, and are also referred to as computer-intensive methods, Davison and Hinkley [6]. These methods allow one to tackle a wide range of problems without having to simplify complex problems, Davison and Hinkley [6]. This approach can be used even in more simple problems to check the adequacy of measures obtained and give approximate solutions, Davison and Hinkley [6]. This is one of the main reasons behind the bootstrapping methods used in this thesis, they provide another measure for the comparison of the Gamma distribution to the lognormal distribution in environmental applications.

The beginning of this Section contains basic definitions and properties, most of which are also presented in Chapter 2, Section 7. These definitions and properties are then followed by derivations of some of the more complex ideas presented in Chapter 2. There are many books and papers written on bootstrapping methods all with different variations of basically the same technique. Throughout this paper I have chosen to use the methods,

notation and terminology of that used by Davison and Hinkley [6]. I've found these methods and applications are well suited for the purposes of this thesis.

Definitions and Properties

Nonparametric means no distributional assumptions are made for the data

Heaveside function $H(u) = \begin{Bmatrix} 0, & u < 0 \\ 1, & u \geq 0 \end{Bmatrix}$, also known as the unit step function

X_i^* represents the ith simulated data set.

N is the number of simulated data sets.

T is the statistic that is used to make inferences on a characteristic of a population.

T_i^* statistic of interest calculated from the ith simulated data set.

t is the value of T from a sample.

Empirical Distribution assigns equal probabilities $1/n$ at each sample value, x_i.

Empirical Distribution Function (EDF) $\hat{F}(x) = \dfrac{\#(x_i \leq x)}{n}$.

$t(F) = \theta$ in nonparametric analysis is the algorithm that defines the parameter of interest, denoted as θ in this instance.

$t = t(\hat{F})$ implies that t is a statistical function of the empirical distribution. function, and the sample estimate of the relationship $t(F) = \theta$.

t(·) is the function for estimating θ.

t represents the scalar estimate of θ.

Examples (Statistical Functions)

The mean $t(F) = \int xdF(x)$ which is estimated from the sample data as

$$\bar{x} = t(\hat{F}) = \int xdF(x) = \int xd\left(1/n\sum H(x-x_i)\right) = 1/n\sum xdH(x-x_i) = 1/n\sum x_i$$

The focus will be on the distributional characteristics of T, for example it's bias and variance, and how to use it to calculate confidence limits for θ.

Basic Confidence Interval Based on Normality and Estimates of Bias and Variance for *T*.

Assume that the distribution of T follows an approximate normal distribution with a mean $\theta + B$ and a variance v, where B and v are the unknown bias and variance. We can construct a confidence interval for θ in the normal way as,

$$P(z_{\alpha/2} < \frac{t-(\theta+B)}{v^{1/2}} < z_{1-\alpha/2}) = 1-\alpha$$

which gives the confidence 100(1-α) % confidence intervals,

$t - B - v^{1/}z_{1-\alpha/2}, t - B - v^{1/2}z_{\alpha/2}$, Davison and Hinkley [6]

Of course the bias and variance are usually not known and can be estimated from there definitions

$B = b(F) = E(T/F) - t(F)$ and $v = v(F) = \text{var}(T/F)$, Davidson and Hinkley [6]

which are estimated from the sample

$B = b(\hat{F}) = E(T/\hat{F}) - t(\hat{F})$ and $v = v(\hat{F}) = \text{var}(T/\hat{F})$, Davison and Hinkley [6]

Moment estimates

Suppose we want to estimate properties of T, say $T-\theta$, from simulated data sets. We would generate simulated data sets by randomly picking data points from the original data set, yielding the simulated data sets, $X_1^*, X_2^*, ..., X_n^*$. For example a bootstrapped estimate for the bias discussed above would be,

$$B = b(\hat{F}) = E(T/\hat{F}) - t(\hat{F}) = E^*(T^*) - t$$

which is then estimated by simulation as,

$$B_N = 1/N\sum_{i=1}^{N} T_i^* - t \text{, } t \text{ estimated from the original data}$$

An estimate for the variance would be,

$$V_N = 1/(N-1)\sum_{i=1}^{N}(T_i^* - \overline{T}^*)^2 \text{ , Davison and Hinkley [6].}$$

These empirical approximations are governed by the law of large numbers and converge to the exact value under the fitted model as N grows large. Davison and Hinkley [6] make an important point in that we are not estimating properties of T in this case but properties of T relative to θ.

Distribution and Quantile estimates

In the preceding discussion we found simulated estimates for the bias and variance of $T-\theta$, but are more commonly based on normal approximations, Davison and Hinkley [6]. However sometimes assuming that T has an approximate normal distribution can be inaccurate. In this paper $Q-Q$ plots, which plot the ordered simulated values

$t_1^*, t_2^*, \ldots, t_N^*$ against expected ordered normal values, are used to measure the appropriateness of using normal approximations.

Without using any approximate distributional model we can look at the quantile estimates for T. Using the above example we can estimate the distribution of $T-\theta$ by the simulated estimate, T^*-t and then estimate the cumulative probabilities by the empirical distribution function of the simulated values t^*-t. Mathematically this is,

$G(c)=P[T-\theta \le c]$ is simulated by $\hat{G}_N(c)=\frac{\#(t_i^*-t \le c)}{N}$, Davison and Hinkley [6].

The basis for this approximation is that $\hat{G}_N(c)$ converges to the exact CDF of T^*-t, as N becomes large. So an α quantile estimate of $T-\theta$ is the $N\alpha$ value of the ordered t^*-t values. Of course it is easier if N is a value such that $N\alpha$ is an integer. Mathematically the α quantile estimate of $T-\theta$ is simply $t_{N\alpha}^*-t$.

Davison and Hinkley [6] note that simulated quantile estimates of this type are in principle better than results based on a normal approximation, because no distributional form of T^*-t is assumed.

Nonparametric Simulation and Exact Moments Under Sampling From the EDF

Again we are assuming no parametric model and that the data are independently and identically distributed from an unknown CDF, F. In nonparametric simulation we are going to use the EDF, $\hat{F}$, as we would a parametric model, Davison and Hinkley [6]. Using this EDF we are going to use simulation of data sets to estimate theoretical properties. To simulate the EDF we assign equal probabilities on the original data values $x_1 x_2, \ldots, x_n$ and then independently sample at random from these data values to yield a

simulated data set X^*. This formally defines the procedure known as the nonparametric bootstrap, Davison and Hinkley [6].

Exact moments under sampling from the EDF are,

Average

$$E^*(\bar{X}^*) = E^*(X^*) = 1/n\sum_{i=1}^{n} x_i = \bar{x}\text{, Davison and Hinkley [6]}$$

Variance of the average

$$\text{var}^*(\bar{X}^*) = 1/n\,\text{var}^*(X^*) = \frac{1}{n}E^*[X^* - E^*(X^*)]^2 = 1/n * \sum_{i=1}^{n} 1/n(x_i - \bar{x})^2$$

$$= \frac{(n-1)}{n} * \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2\text{, Davison and Hinkley [6]}$$

or

$$\text{var}^*(\bar{X}^*) = s^2/n^2$$

This derivation given by Davison and Hinkley [6] can also be explained using the following corollary to a theorem from Hogg and Craig [7], pg.220.

> Corollary: Let $X_1, X_2, ..., X_n$ denote observations of a random sample of size n from a distribution that has mean μ and variance σ^2. The mean and variance of $Y = \sum_{i=1}^{n} k_i X_i$ are respectively, $\mu_y = \left(\sum_{i=1}^{n} k_i\right)\mu$ and $\sigma_y{}^2 = \left(\sum_{i=1}^{n} k_i{}^2\right)\sigma^2$.

To use this theorem note that we are sampling from the EDF $\hat{F}$ that has a mean $\bar{x}$ and variance $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$. Therefore the simulated data set, say Y, is made up of independently sampled X_i's. So applying the corollary,

$$\overline{Y} = \sum_{i=1}^{n} 1/nX_i \text{ , each } k_i = 1/n, \Rightarrow \mu_y = \left(\sum_{i=1}^{n} 1/n\right)\overline{x}$$

and variance,

$$\sigma_y{}^2 = \left(\sum_{i=1}^{n} (1/n)^2\right) s^2 = 1/n\left(\sum_{i=1}^{n} (x_i - \overline{x})^2 / n\right) //$$

Nonparametric Bootstrapping Basic Confidence Intervals

The first of the basic bootstrapped confidence intervals are based on the quantile estimates discussed above which are derived from the following probability statement,

$$P(a \le T - \theta \le b) = 1 - \alpha \Rightarrow P(b \le \theta \le a) = 1 - \alpha$$

Substituting the quantile estimates for $T - \theta$ yields the basic bootstrap confidence interval,

$$t - (t^*_{N(1-\alpha/2)} - t), t - (t^*_{N(\alpha/2)} - t)\text{, Davison and Hinkley [6]}$$

The next basic confidence interval is based on mimicking the studentized t statistic which is of the form,

$$S = \frac{T - \theta}{v^{1/2}}$$

which is estimated by the studentized bootstrap statistic,

$$S^* = \frac{(T^* - t)}{V^{*1/2}} \text{ with simulated values } s_i^* = \left(\frac{t_i^* - t}{v_i^{*1/2}}\right)\text{, Davison and Hinkley [6].}$$

The delta method variance, discussed next, is substituted into these equations since we are assuming no parametric model which gives the studentized bootstrap statistic based on nonparametric resampling as,

$$S = \frac{T-\theta}{v_L^{1/2}} \text{ with simulated values } s_i^* = \left(\frac{t_i^* - t}{v_{Li}^{*1/2}} \right)$$

The $100(1-\alpha)\%$ studentized bootstrap confidence limits, also referred as bootstrap-t limits, are generated by substituting the simulated quantile estimates of S into the well known studentized confidence intervals yielding,

$(t - v_L^{1/2} s^*_{N(1-\alpha/2)},\ t - v_L^{1/2} s^*_{N\alpha/2})$, Davison and Hinkley [6]

The above methods are simplified by using a normal approximation for the estimated quantiles. Substituting into the well-known normal confidence intervals yields the $100(1-\alpha)\%$ confidence interval,

$t \mp v^{1/2} z_{1-\alpha/2}$ which, for the nonparametric case is $t \mp v_L^{1/2} z_{1-\alpha/2}$

Davison and Hinkley [6] suggests that if the distribution of T - θ is dependent on unknowns alternative expressions contrasting T and θ, such as the studentized comparisons should be used. Moreover, the first two types should be used when a normal approximation to the quantiles is inadequate. The inadequacy of a normal approximation for this thesis is determined by viewing the normal Q-Q plot of the simulated values, t^*.

Theory Behind Delta Methods

Following is a discussion of the theory behind delta methods that will make it easier to understand the concepts of the influence function and the nonparametric delta method.

When doing a parametric analysis (assuming a distributional form) often it is possible to represent estimators, T, in terms of fundamental statistics, $U_1,...U_n$, for which exact or approximate distributional calculations are possible, Davison and Hinkley [6]. Moment

estimates would be an example. If this is the case the delta methods can be used to find distributional approximations of T itself, Davison and Hinkley [6].

To show this, consider the case where T is a scalar estimator that is a function of the scalar statistic U based on a sample of size n, $T = g(U)$. Suppose the following,

$$(1)\ U \approx N(\xi, \sigma_\xi{}^2 / n) \Rightarrow (2)\ Z = \frac{U - \xi}{\sigma_\xi / \sqrt{n}} \sim N(0,1)$$

therefore we can write U as,

$U = \xi + \sigma_\xi / nZ + O_p(n^{-1})$, where $O_p(n^{-1})$ is the hidden bias of T, since (1) is only approximately normal. Also we can write, $U = \xi + o_p(1)$ which is a statement of the consistency property of U. Next consider $T = g(U)$, where g is a smooth differentiable function. Provided that $g'(\xi) \neq 0$, then,

$$(3)\ T \approx N\left(\theta, \frac{g'(\xi)^2 \sigma_\xi{}^2}{n}\right), \text{ this will be shown below}$$

This result is usually called the delta method result, Davison and Hinkley [6].

The main feature of this result is the delta method variance approximation defined

(4) $\text{var}[g(U)] \cong [g'(\xi)]^2\ \text{var}(U)$, Davison and Hinkley [6].

This is derived as follows,

Note if g is smooth then T is consistent for $\theta = g(\xi)$, since from (1),

$g(U) = g(\xi + o_p(1)) = g(\xi) + o_p(1)$, Davison and Hinkley [6]

Using Taylor series expansions to find a polynomial that best estimates $g(U)$ near $U = \xi$, we have,

$$T = g(U) = g(\xi) + g'(\xi)(U - \xi) + \frac{g''(\xi)}{2}(U - \xi)^2 + o_p(n^{-1}),\ \text{Hinkley et al [6]}$$

since the remainder is proportional to $(U-\xi)^3$, the next term in the Taylor series expansion, a truncated version is,

$$T = g(U) = g(\xi) + g'(\xi)(U-\xi) + o_p(n^{-1/2}),$$ Davison and Hinkley [6]

Using (2) we have,

$$Z = \frac{U-\xi}{\sigma_\xi / \sqrt{n}} \sim N(0,1) \Rightarrow (U-\xi) = \sigma_\xi / \sqrt{n}$$, thus

$$T = g(U) = g(\xi) + g'(\xi)(U-\xi) + o_p(n^{-1/2}) = g(\xi) + \frac{g'(\xi)\sigma_\xi Z}{\sqrt{n}} + o_p(n^{-1/2})$$

which implies, since T is a linear function of Z, which is normal, that (3) is true,

$$(3)\quad T \approx N\left(\theta, \frac{g'(\xi)^2 \sigma_\xi{}^2}{n}\right)$$

and from (1) we have $\text{var}(U) = \sigma_\xi{}^2 / n$, therefore,

$$\text{var}[g(U)] = g'(\xi)^2 \frac{\sigma_\xi{}^2}{n} = g'(\xi)^2 \text{ var}(U)$$, which is equation (4). Davison and Hinkley [6].

The bias of T, that is hidden in the $o_p(n^{-1})$ term of the Taylor expanded series above, can be estimated by taking the expectation ignoring the remainder term,

$$T = g(U) = g(\xi) + g'(\xi)(U-\xi) + \frac{g''(\xi)}{2}(U-\xi)^2$$

which is,

$$E(T) = E(g(U)) = g(\xi) + g'(\xi)E((U-\xi)) + \frac{g''(\xi)}{2}E((U-\xi)^2)$$

$$= g(\xi) + g'(\xi)(\xi - \xi) + \frac{g''(\xi)}{2}\sigma_{\xi}^{2}/n, \text{ since } E(U-\xi)^2 = \text{var}(U) = \sigma_{\xi}^{2}/n.$$

If U is unbiased for ξ, this simplifies to

$$E(T) \cong \theta + \frac{1}{2n}g''(\xi)\sigma_{\xi}^{2}, \text{ Davison and Hinkley [6].}$$

As a lead in to the next section Davison and Hinkley [6] state the above results can be extended to the case when U is a set of observed frequencies $f_1,\ldots,f_n$ when the random variable is discrete with probabilities $p_1,\ldots,p_n$ on n possible values. In this case the analogue of $T = g(U) = g(\xi) + g'(\xi)(U-\xi) + \frac{g''(\xi)}{2}(U-\xi)^2 + o_p(n^{-1})$ is,

$$(5)\ T \cong g(p_1,\ldots p_n) + \sum_{i=1}^{n}\left(\frac{f_i}{n} - p_i\right)\frac{\partial(p_1,\ldots,p_n)}{\partial p_i}, \text{ Davison and Hinkley [6].}$$

Nonparametric Delta Method and the Influence Function

The influence function is an extension of (5) using the Taylor series expansion to statistical functions, Davison and Hinkley [6]. The linear form of the expansion of (5) is,

$$(6)\ t(G) \cong t(F) + \int L_t(y;F)dG(y), \text{ where } L_t \text{ is the first derivative of t } (\bullet) \text{ at } F,$$

defined as,

$$L_t(y;F) = \lim e \to 0 \frac{t[(1-e)F + eH_y] - t(F)}{e} = \frac{\partial[(1-e)F + eH_y}{\partial e}\Big|_{e=0}$$

where $H_y(u) \equiv H(u-y)$, is the Heaveside function or unit step function jumping from 0 to 1 at $u = y$, Davison and Hinkley [6].

The resulting definitions are,

influence function- $L_t(y) = L_t(y;F)$

empirical influence function- $l(y) = L_t(y; \hat{F})$ (estimated from sample, $\hat{F}$)

empirical influence values- $l_i = l(y_i)$, Davison and Hinkley [6].

Applying equation (6) with $G = \hat{F}$ gives the nonparametric delta method,

$$(7)\ t(\hat{F}) \cong t(F) + \int L_t(y; F)d\hat{F}(y) = t(F) + \frac{1}{n}\sum_{i=1}^{n} L_t(y_i; F)$$

where the right hand side is known as the linear approximation, Davison and Hinkley [6].

Next the central limit theorem is applied to the sum $\frac{1}{n}\sum_{i=1}^{n} L_t(y_i; F)$ yielding,

$$T - \theta \approx N(0, v_L(F))$$

because $\int L_t(y; F)dF(y) = 0$, and where

$v_L(F) = n^{-1}\ \mathrm{var}[L_t(Y)] = n^{-1}\int L_t^2(y)dF(y)$, Davison and Hinkley [6].

Davison and Hinkley [6] note that in practice $v_L(F)$ is approximated by substituting the empirical distribution function, $\hat{F}$ for F, obtaining the nonparametric delta method variance estimate,

$$v_L = v_L(\hat{F}) = n^{-2}\sum_{i=1}^{n} l_i^2$$

Davison and Hinkley [6] note that equation (7) implies that,

$$\int L_t(y; \hat{F})d\hat{F}(y) = n^{-1}\sum_{i=1}^{n} l_i \equiv 0$$

In cases when then derivative , $L_t(y; F) = \lim e \to 0 = \frac{\partial[(1-e)F + eH_y}{\partial e}\Big|_{e=0}$ is difficult to evaluate theoretically, the numerical approximation can be used,

$$L_t(y; F) \cong \frac{t[(1-e)F + eH_y] - t(F)}{e}$$ with a small value of e, like $1/(100n)$.

Davison and Hinkley [6] provide the following example, which is used in this paper.

Ex: Let $t = \bar{y}$, which has the statistical function $t(F) = \int y dF(y)$. Then the influence function,

$$L_t(y;F) = \lim e \to 0 \frac{t[(1-e)F + eH_y] - t(F)}{e} = \frac{\partial[(1-e)F + eH_y}{\partial e} \Big|_{e=0}$$

is evaluated by setting $t[(1-e)F + eH_y] = (1-e)\mu + ey$ and differentiated to obtain

$$L_t(y) = \frac{\partial[(1-e)\mu + ey]}{\partial e} \Big|_{e=0} = y - \mu$$, Davison and Hinkley [6]. $l(y) = y - \bar{y}$

the results are,

empirical influence function- $l(y) = y - \bar{y}$

empirical influence values- $l_i = y_i - \bar{y}$

therefore the delta method variance approximation is,

$v_L = (n-1)s^2 / n^2$, where $s^2 = \sum_{i=1}^{n} (y - \bar{y})^2 / (n-1)$, Davison and Hinkley [6].

This is arrived at as follows,

$$v_L = v_L(\hat{F}) = n^{-2} \sum_{i=1}^{n} l_i^2 = n^{-2} \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)n^{-2}s^2 .$$

References

[1] Choi, S.C., Wette, R. Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias. Technometrics, v. 11, No. 4, 683. (1969)

[2] Berkey, Dennis D., Calculus Second Edition. Saunders College Publishing, New York, N.Y., 1988

[3] Burden, Richard L., Douglas, J. Faires., Numerical Analysis Sixth Edition, Brooks/Cole Publishing Company, Pacific Grove, C.A., 1997

[4] Johnson, N.L., Kotz, S., Continuous Univariate Distributions. Houghton-Mifflin, Boston, Mass, 1970

[5] Schneider, Bruce E. Kolmogrov-Smirnov Test Statistics for the Gamma For the Gamma Distribution With Unknown Parameters. Dissertation, Temple University, 1978

[6] Grice, John V., Bain, Lee J. Inferences Concerning the Mean of the Gamma Distribution. Journal of the American Statistical Association, Dec., v. 75, No. 372, 929. (1980)

[7] Hogg, Robert V., Craig, Allen T., Introduction to Mathematical Statistics. Prentice-Hall, Englewood Cliffs, N.J., 1995

[8] Mood, A.M., Graybill, F.A., Boes, D.C., Introduction to the Theory of Statistics McGraw Hill, New York, NY, 1974

[9] Linhart, H. Approximate Confidence Limits For The Coefficient Of Variation Of Gamma Distributions. Biometrics, Sept., 733. (1965)

[10] Bartlett, M.S., Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society, A.160, 268-282, 1937

[11] Pairman, E., Tables of the Digamma and Trigamma Function. Cambridge University Press, Cambridge, England, 1954

[12] Stephens, M.A., Use of the Kolomgrov-Smirnov, Cramer-von Mises and Related Statistics Without Extensive Tables. Journal of the Royal Statistical Society, Series B. 31, 115-122, 1970

[13] Durbin, J., Kolmogrov-Smirnov Tests When Parameters Are Estimated With Applications to Tests of Exponentiality and Tests on Spacings. Biometrika, 62, 5-22, 1975

[14] Whittaker, J. Generating Gamma and Beta Random Variables with Non-integral Shape Parameters. Appl. Statist, 23, No. 2, 210. (1974)

[15] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes in Fortran, Cambridge University Press, 1994

[16] Bradu, D., Mundlak, Y., Estimation in Lognormal Linear Models. Journal of the American Statistical Society 65 (198-211), 1970

[17] Finney, D.J., " On the Distribution of a Variate whose Logarithmic is Normally Distributed," Supplement to the Journal of the Royal Statistical Society, Vol 7, p. 155-161, 1951

[18] Land, C.E., Confidence Intervals for Linear Functions of the Normal Mean and Variance. Annals of Mathematical Statistics, 42, 11187-1205, 1971

[19] Land, C.E., 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, 1975

[20] Gilbert, Richard O., " Statistical Methods For Environmental Pollution Monitoring,"Van Nostrand Rheinhold, New York, NY, 1987

[21] Gerald, C., Wheatley, P.O., Applied Numerical Analysis. Addison-Wesley Publishing Company, Reading, Mass. 1970

[22] Davison, A.C., Hinkley, D.V. Bootstrap Methods and their application. Cambridge University Press, New York, NY, 1997.

CHAPTER 4

FORTRAN SOURCE CODE

Provided below is the full Fortran 77 code used to obtain the results presented in Chapter 2. When applicable this program was tested against results obtained by other authors. One example is given in the results section of Chapter 2 and another is found in Chapter 2, Section 7.

```
Program Gamma
Parameter definitions. All real numbers throughout the program are in double Precision.

      Double Precision a(10000),z(10000),xbar,g(10000),alpha,beta,h(12)
      Double Precision df,chi,ub,n1,lngamma,mu,x,prg,u(10000)
      Double Precision gd(10000),avg(10000),avgd(10000)
      Double Precision lnormd(10000),plnormd(10000),pi
      Double Precision ltran(10000),stdlt,ltbar
      Double Precision muln,medln,varln,stdln,mlemu,mlestd
      Double Precision ltcv,mlecv,errmln,mlesd,mleupper,cheby
      Double Precision std1,std2,cltucl,cltadjucl,cheby1
      Double Precision d(30000),bia,vr,bsucl,studl,studl1,stbl
      Double Precision stbl1,pucl,prz(1000),kst,kst1,kst2,zm
      Double Precision alpha4,alpha3,gdev(100000),lngamma3,lngamma4
      Double Precision tp,pval,tbcaper(4),tbcanp(4),hucl,hh
      Double Precision lnxb,lnsd
      Integer i,j,n,ct,n2,m1,e,n3
c ********************** PI*************************************************
      pi=3.141592653589793
c ******************************************************************************
      write(*,*)'Enter the number of data points'
      Read *, n
      write(*,*)'Enter number of data sets parametric bootstraping'
      Read *,n2
      write(*,*)'Enter number of data sets non-par bootstrapping'
      Read *,n3
      write(*,*)'Enter H statistic'
```

```
      Read *,hh
      write(*,*)'Enter lognormal distribution mean'
      Read *,lnxb
      write(*,*)'Enter lognormal distribution standard deviation'
      Read *,lnsd
      write(*,*)' ',lnxb,lnsd
      n1=n
      x=2.0
      m1=int(.95*n2)
      e=1000*n

c ********************** DATA FILE RETRIEVAL ******************************
```

The first file to be opened and read is the data set, followed by the file 'dataa' which is a table of the hermite polynomials used in the subroutine confidencebound to provide a 95% upper bound for the mean assuming a Gamma distribution. The last file is a random set of integers from 1 to n used in the nonparamteric bootstrap routines. This fixed data set was used to provide the same bootstrapped values during the writing phase and can easily be replaced with a random generating function.

```
      open (unit=15, file='ln43',status='old')
      do j=1,n
        read(15,*)a(j)
      enddo
      close(15)
      open ( unit=16, file='dataa',status='old')
      do j=1,12
        read(16,*) h(j)
      enddo
      close(16)
      open (unit=15, file='data15',status='old')
      do j=1,e
        read(15,*)d(j)
      enddo
      close(15)

c ***************** MAIN ********************************************************
      call sort(a,n)
      call basicstat(n,a,std1,std2,cltucl,cltadjucl,cheby1)
      call newton(a,n,xbar,beta,alpha,zm)
      call confidencebound(alpha,n1,xbar,h,df,chi,ub)
      call scale(a,beta,n,z)
      call gser(alpha,z,n,prg,lngamma,prz)
      call kstat(n,prz,kst)

c ****************** ALPHA3,4 ***************************************************
```

The values alpha3 and alpha4 are the upper and lower confidence limits discussed in Chapter 2, Section3, and are input here, if desired, to calculate the Kolmogrov-Smirnov test statistic of Chapter 2,Section 4.

```
      write(*,*)'enter alpha upper CI'
```

```
      read *,alpha3
      write(*,*)'enter alpha lower CI'
      read *,alpha4
      call gser(alpha3,z,n,prg,lngamma3,prz)
      call kstat(n,prz,kst1)
      call gser(alpha4,z,n,prg,lngamma4,prz)
      call kstat(n,prz,kst2)

c ****************** LOGNORMAL CALLS ************************************
      call ltstat(n,a,ltran,ltbar,stdlt,muln,medln,varln,stdln,mlemu
     c      ,mlestd,ltcv,mlecv,errmln,mlesd,mleupper,cheby,hucl
     c      ,hh)
```

The bootstrapping parametric model to be used needs to be determined and the correct subroutine called. For example if a Gamma model is the assumed model then the call of the lognormal subroutine needs to be commented out and vice versa. The generated data set is sent to pgboot as gdev.

```
      call nonparbootstrap(n,n3,a,d,xbar,std1,bia,vr,bsucl,studl
     c     ,studl1,stbl,stbl1,pucl,tbcanp)
c     call gammadeviates(n,n2,alpha,beta,gdev)
      call gammadeviates(n,n2,alpha,beta,gdev)
      call lognormal(n2,gdev,lnsd,lnxb)
      call pgboot(n,n2,alpha,xbar,ltbar,mlesd,lngamma,pi,gdev,tp,pval
     c ,tbcaper)
c ********************* END MAIN ********************************************

c ******************** OUTPUTS ********************************************
      write(*,*)'Results from data'
      write(*,*),'xbar                    ',xbar
      write(*,*),'standard deviation unbiased  ',std1
      write(*,*),'standard deviation mle       ',std2
      write(*,*),'CLT 95% upper bound          ',cltucl
      write(*,*),'CLT adjusted 95% upper bound ',cltadjucl
      write(*,*),'chebychev 95% upper bound    ',cheby1
      write(*,*)' '
      write(*,*)'Gamma pdf results'
      write(*,*),'alpha                   ',alpha
      write(*,*),'beta                    ',beta
      write(*,*),'zm                      ',zm
      write(*,*),'2nk^                    ',df
      write(*,*),'Chi 2nk^                  ',chi
      write(*,*),'95% upper bound             ',ub
      write(*,*),'k-smir alpha              ',kst
      write(*,*),'k-smir alpha upper CI limit  ',kst1
      write(*,*),'k-smir alpha lower CI limit  ',kst2
      write(*,*),'bootstrap p-val             ',pval
      write(*,*),'bca perentile 95% ub         ',tbcaper(4)
      write(*,*),' '
      write(*,*)' '
```

```
      write(*,*)'log transformed data results'
      write(*,*),'mean                    ',ltbar
      write(*,*),'standard deviation      ',stdlt
      write(*,*),'MLE standard deviation  ',mlesd
      write(*,*),'coefficient of variance ',ltcv
      write(*,*),'H Stat UCL              ',hucl
      write(*,*)' '
      write(*,*)'results for ln pdf using MLEs'
      write(*,*),'mean                    ',mlemu
      write(*,*),'standard deviation      ',mlestd
      write(*,*),'coefficient of variance ',mlecv
      write(*,*),'95% UCL                 ',mleupper
      write(*,*)' '
      write(*,*)'results for ln pdf using MVUE theory'
      write(*,*),'mean                    ',muln
      write(*,*),'median                  ',medln
      write(*,*),'standard deviation      ',stdln
      write(*,*),'std error of mean       ',errmln
      write(*,*),'chebychev 95% UCL       ',cheby
      write(*,*)' '
      write(*,*)'Non Parametric bootstrapping results'
      write(*,*),'bias est Br             ',bia
      write(*,*),'var  est Vr             ',vr
      write(*,*),'basic                   ',bsucl
      write(*,*),'studentized             ',studl
      write(*,*),'studentized vl          ',studll
      write(*,*),'standard w normal       ',stbl
      write(*,*),'standard w normal vl    ',stbll
      write(*,*),'basic percentile UCL    ',pucl
      write(*,*),'bca perentile 95% ub    ',tbcanp(4)
      write(*,*)' '
      END
```

c ******* FUNCTIONS **

c ******* GMT ***

Returns the value of Finney function,

$$g_n(t) = \sum_{k=0}^{\infty} w_k(t) = \sum_{k=0}^{\infty} \frac{n^k(n+2k)}{n(n+2)...(n+2k)} \left(\frac{n}{n+1}\right)^k \frac{1}{k!} t^k$$

used in the MVUE estimates for a lognormal distribution, presented in Chapter 2, Section 6. The function is stopped after the previous and current value of the function have an absoloute difference less than the tolerance, tol=0.00000000000001.

```
      Function gmt(n,stdl,a)
      Double Precision x0,x1,x3,x4,m,stdl,a,itmax,tol
      Integer p,n
      itmax=1.0
      tol=0.00000000000001
      m=n
```

```
      x0=0.0
      x4=1.0
      p=0
      do while(itmax.gt.tol)
       x3=((m**p)*(m+(2*p)))/(x4*(m+(2*p)))
       x4=1/(x3*(1/((m**p)*(m+(2*p)))))
       x1=x0+(((a*((stdi)**2))**p)*(1/(factrl(p)))
     c    *((m/(m+1))**p)*x3)
       itmax=abs((x1-x0))
       x0=x1
       p=p+1
      enddo
      gmt=x1
      return
      END
```

c ******* PRBLN**********

Returns the probability of a random variable from a lognormal distribution. Used in the nonparametric bootstrap percentile estimates given in Chapter 2, Section 7.

```
      Function prbln(x)
      Double Precision x,x1,x2,x3
      Parameter (d1=.0498673470d0,d2=.0211410061d0,d3=.0032776263d0,
     c          d4=.0000380036d0,d5=.0000488906d0,d6=.0000053830d0)
      x1=(d1*x)+(d2*(x**2))+(d3*(x**3))
      x2=(d4*(x**4))+(d5*(x**5))+(d6*(x**6))
      x3=1-((1/((1+x1+x2)**16))*0.5)
      prbln=x3
      return
      END
```

c ******* ZINPRB**********

Returns the inverse probability of a normal distribution. Used in the nonparametric bootstrap percentile estimates given in Chapter 2, Section 7.

```
      Function zinprb(xx)
      Double Precision x,xx,c0,c1,c2,d1,d2,d3,t
      Parameter (c0=2.515517d0,c1=.802853d0,c2=.010328d0,d1=1.432788d0,
     c          d2=.189269d0,d3=.001308d0)
      if (xx.ge.0.5)then
        xx=1.0-xx
      t=SQRT(log((1/xx)**2))
      x=t-((c0+(c1*t)+(c2*(t**2)))/(1+(d1*t)+(d2*(t**2))+(d3*(t**3))))
      zinprb=x
      else
      t=SQRT(log((1/xx)**2))
      x=t-((c0+(c1*t)+(c2*(t**2)))/(1+(d1*t)+(d2*(t**2))+(d3*(t**3))))
      zinprb=-x
      endif
      return
```

```
      END

c ******* FACTORIAL ************************************************************
```

Returns n! using the function gammln.

```
      Function factrl(m)
      Integer m,j,ntop
      Real a(33),gammln
      Save ntop,a
      Data ntop,a(1)/0,1.0/
      if (m.lt.0) then
        pause'negative factorial'
      else if (m.le.ntop) then
        factrl=a(m+1)
      else if (m.le.32) then
        do j=ntop+1,m
          a(j+1)=j*a(j)
        enddo
        ntop=m
        factrl=a(m+1)
      else
        factrl=exp(gammln(m+1.0))
      endif
      return
      END

c ******* GAMMLN ***************************************************************
```

Returns the value of the natural logarithm of the Gamma Function, $\ln\Gamma(x)$.

```
      Function gammln(xx)
      Real gammln
      Integer j
      Double Precision ser,stp,tmp,x,y,cof(6),xx
      Save cof,stp
      Data cof,stp/76.18009172947146d0,-86.50532032941677d0,
     c  24.01409824083091d0,-1.231739572450155d0,.1208650973866179d-2,
     c  -.5395239384953d-5,2.5066282746310005d0/
      x=xx
      y=x
      tmp=x+5.5d0
      tmp=(x+0.5d0)*log(tmp)-tmp
      ser=1.000000000190015d0
      do j=1,6
        y=y+1.d0
        ser=ser+cof(j)/y
      enddo
      gammln=tmp+log(stp*ser/x)
      return
      END
```

c ******* NORMDEV **

Returns a single random variable from a standard normal distribution.

```
      Function normdev(xx)
      Integer xx,set
      Real ran1
      Double Precision fac,gset,rsq,v1,v2,normdev
      Save iset,gset
      Data iset/0/
      if (iset.eq.0) then
1       v1=2.0*ran1(xx)-1.0
        v2=2.0*ran1(xx)-1.0
        rsq=v1**2+v2**2
        if(rsq.ge.1.0.or.rsq.eq.0.0) goto 1
        fac=SQRT(-2.0*log(rsq)/rsq)
        gset=v1*fac
        normdev=v2*fac
        iset=1
      else
        normdev=gset
        iset=0
      endif
      return
      END
```

c ******* RAN1 **

Returns a random deviate x, $0 < x < 1$, shuffling the output to remove low-order serial correlations.

```
      Function ran1(idum)
      Double Precision AM,EPS,RNMX
      Real ran1
      Integer idum,IA,IM,IQ,IR,NTAB,NDIV
      Parameter (IA=16807,IM=2147483647,AM=1.0/IM,IQ=127773,IR=2836,
     c     NTAB=32,NDIV=1+(IM-1)/NTAB,EPS=1.2e-7,RNMX=1.-EPS)
      Integerj,k,iv,iv(NTAB),iy
      Save iv,iy
      Data iv /NTAB*0/, iy /0/
      if (idum.le.0.or.iy.eq.0) then
        idum=max(-idum,1)
        do j=NTAB+8,1,-1
          k=idum/IQ
          idum=IA*(idum-k*IQ)-IR*k
          if (idum.lt.0) idum=idum+IM
          if(j.le.NTAB) iv(j)=idum
        enddo
      endif
      k=idum/IQ
      idum=IA*(idum-k*IQ)-IR*k
      if (idum.lt.0) idum=idum+IM
```

```
      j=1+iy/NDIV
      iy=iv(j)
      iv(j)=idum
      ran1=min(AM*iy,RNMX)
      return
      END

c ******* SUBROUTINES ***************************************************

c ******* NONPARBOOTSTRAP ************************************************
```

This subroutine calculates all the bootsrapped upper bounds and related values assuming no parametric model presented in Chapter 2, Section 7.

```
      Subroutine nonparbootstrap(m,m1,a0,dd,est,std,bias,sd,basucl
     c    ,studucl,studucl1,stbucl,stbucl1,perucl,tbca)
      Double Precision a0(m),aa(23000),a1(23000),dd(30000)
      Double Precision xbb(m1),x1,xbt,b(m1),s(m1),sd
      Double Precision est,stbucl,x2,ss(23000),sd1(1000)
      Double Precision t(m1),pivucl,std,t1(m1),tl(m1),sdvlyb
      Double Precision basucl,studucl,bias,vlyy,vlyb,stbucl1
      Double Precision lg(m1),studucl1,sdvl(m1),vlcon,vlcon1,a,bc
      Double Precision perucl,x5,w,h,cc(4),ck(4),ck1(4),tbca(4)
      Double Precision alph(4),si(4),zh(4),ah(4),m2,r(4),rrd(4)
      Integer j,m,n,n1,p1,p2,h2,rr(4)
      m2=m1
      n=1
      n1=m
      p1=.05*m1
      p2=.95*m1

c ********************** bootstrap routine **********************
      do i=1,m1
      x1=0.0
      do j=n,n1
         aa(j)=a0(dd(j))
         a1(j)=x1+aa(j)
         x1=a1(j)
      enddo
      xbb(i)=x1/m

      x2=0.0
      do j=n,n1
        ss(j)=x2+((aa(j)-xbb(i))**2)
        x2=ss(j)
      enddo
      sd1(i)=SQRT((x2/m-1))

       n=n+m
       n1=n1+m
      enddo
```

```
c ********** calc vl(i) for studentized z ************************
      vlcon=m-1
      vlcon1=(sqrt(vlcon))/m
      do j=1,m1
        sdvl(j)=vlcon1*sd1(j)
      enddo
c
*****************************************************************************
     x1=0.0
      do j=1,m1
        b(j)=x1+xbb(j)
        x1=b(j)
      enddo
      xbt=x1/m1
      bias=xbt-est

      x1=0.0
      do j=1,m1
        s(j)=x1+((xbb(j)-xbt)**2)
        x1=s(j)
      enddo
      sd=SQRT(x1/(m1-1))

c *********** log of t's for normal check*****************************
Values used to check for normality as in examples 3 and 4 of Chapter 2, see graphs.
      do j=1,m1
        lg(j)=log(xbb(j))
      enddo

c ********** t1 quantile estimates basic boot ************************
c      do j=1,m1
c        t1(j)=xbb(j)-est
c      enddo
      call sort(t1,m1)
      basucl=est-(t1(p1))

c *************** delta method var est ybar ************************
      x1=0.0
      do j=1,m
        vlyy=x1+((a0(j)-est)**2)
        x1=vlyy
      enddo
      vlyb=x1/(m**2)
      sdvlyb=SQRT(vlyb)

c ********************** studentized ***********************************
      do i=1,m1
        t(i)=(xbb(i)-est)/sd1(i)
      enddo
      call sort(t,m1)
      do j=1,m1
```

```
        tl(j)=(xbb(j)-est)/sdvl(j)
      enddo
      call sort(tl,m1)
      studucl=est-(t(p1)*std)
      studucl1=est-(tl(p1)*sdvlyb)

c ********************** Normal approxs ****************************************
      stbucl=est-bias+(1.6449*sd)
      stbucl1=est+(1.6449*sdvlyb)

c ******************** percentile methods ***********************************
      call sort(xbb,m1)
      perucl=xbb(p2)
      x1=0.0
      do j=1,m
        bc=x1+((a0(j)-est)**3)
        x1=bc
      enddo
      a=(bc/(SQRT((vlyy**3))))*(1.0/6.0)
     x5=xbb(1)
     h2=1
     do while(est.gt.x5)
       h2=h2+1
       x5=xbb(h2)
     enddo
     h=h2
     h=h/m2
     w=zinprb(h)
     alph(1)=.025
     alph(2)=.975
     alph(3)=.050
     alph(4)=.950
     si(1)=-1.96
     si(2)=1.96
     si(3)=-1.6449
     si(4)=1.6449
     do j=1,4
       zh(j)=w+si(j)
     enddo
     do j=1,4
       ah(j)=prbln(w+(zh(j)/(1.0-(a*zh(j)))))
     enddo
     do j=1,4
       r(j)=m2*ah(j)
     enddo
     do j=1,4
       rr(j)=int(r(j))
       rrd(j)=rr(j)
     enddo
     do j=1,4
       cc(j)=w+(zh(j)/(1.0-(a*zh(j))))
```

```
 enddo
 do j=1,4
   ck(j)=zinprb(rrd(j)/m2)
 enddo
 do j=1,4
   ck1(j)=zinprb((rrd(j)+1.0)/m2)
 enddo
 do j=1,4
  tbca(j)=xbb(rr(j))+
c ((cc(j)-ck(j))/(ck1(j)-ck(j)))*(xbb(rr(j)+1)-xbb(rr(j)))
 enddo
 return
 END
```

c ******* BASICSTAT **********

Returns all the general statistics of the sample including the upper bounds for the mean based on the Central Limit Theorem, Adjusted Central Limit Theorem and the Chebychev Theorem.

```
 Subroutine Basicstat(m,a0,s1,s2,clt,cltad,chby1)
 Double Precision a0(m),s1,s2,x,x1,xb,x2,x3,clt,l
 Double Precision x4,x5,kk3,zad,z1,z2,cltad,chby1
 Integer m
 l=m
 x1=0.0
 do j=1,m
   x=x1+a0(j)
   x1=x
 enddo
 xb=x1/l
 x3=0.0
 do j=1,m
   x2=x3+((a0(j)-xb)**2)
   x3=x2
 enddo
 s1=SQRT(x3/(m-1))
 s2=SQRT(x3/m)
 clt=xb+(1.6449*(s1/SQRT(l)))
 x3=0.0
 do j=1,m
   x3=x4+((a0(j)-xb)**3)
   x4=x3
 enddo
 kk3=(1/(m*(s1**3)))*x4
 z1=(kk3/(6*SQRT(l)))
 z2=1+(2*(1.6449**2))
 zad=1.6449+(z1*z2)
 cltad=xb+(zad*(s1/SQRT(l)))
 chby1=xb+(4.47*(s1/SQRT(l)))
 return
```

```
      END

c *******PGBOOT ************************************************************
```
Returns all the bootstrapped estimates assuming a parametric model discussed in Chapter 2, Section 8.

```
      Subroutine pgboot(m,m1,k,yb,ltb,mld,lg,pi,gdv,t,p,tbca)
      Double Precision k,k1(m1),yb,ltb,mld,lg,pi,t,p
      Double Precision lg1(m),t1(m1),ltb1(m1),m2
      Double Precision gdv(100000),x1,aa(100000),a1(100000)
      Double Precision xbb(m1),xbt(m1),xb,x2,lb(100000),a2
      Double Precision x,w,wbar,a,a0,m2,psi,dpsi,b,xb,w1
      Double Precision x3,ml(100000),mld1(m1),x4,h1,ww,x5,h3
      Double Precision zh(4),ah(4),r(4),tbca(4),si(4)
      Double Precision alph(4),rrd(4),cc(4),ck(4),ck1(4)
      Integer i,j,m,m1,n,n1,l,h,h2,rr(5)
      n=1
      n1=m
```
Test statistic t based on the original sample used to determine the bootstrapped p-value
```
      t=-k*(log(k/yb))-(k*ltb)+k-(0.5*(log(2.0*pi*(mld**4))))
     c     -0.5+lg

c *********** boot routine calculates xbar alpha beta ***********************************
      do l=1,m1

       x1=0.0
       x2=0.0
       x3=0.0
       w1=0.0

       do j=n,n1
         aa(j)=gdv(j)
         a1(j)=x1+aa(j)
         x1=a1(j)
         w=w1+log(aa(j))
         w1=w
         lb(j)=x2+log(aa(j))
         x2=lb(j)
       enddo
       xbb(l)=x1/m
       ltb1(l)=x2/m

       do j=n,n1
         ml(j)=x3+((log(aa(j))-ltb1(l))**2)
         x3=ml(j)
       enddo
       mld1(l)=x3/m

       xb=x1/m
       wbar=w1/m
```

```
      m2=log(xb)-wbar
      a0=1/(2*m2)
      do i=1,100
        psi=log(a0)-(1+(1-(.1-1/(21*a0))/(a0*a0))/(6*a0))/(2*a0)
        dpsi= (1+(1+(1-((.2-1/(7*a0))/(a0*a0)))/(3*a0))/(2*a0))/a0
        a=a0-((log(a0) - psi - m2)/(1/a0 - dpsi))
        a0=a
      enddo
      if(a0.ge.1.0)then
        a0=(a0*(1-(3/m)))+(2/(3*m))
      endif
      b=xb/a0
      k1(l)=a0
      lg1(l)=gammln(a0)
        t1(l)=-k1(l)*(log(k1(l)/xbb(l)))-(k1(l)*ltb1(l))+k1(l)
     c  -(0.5*(log(2.0*pi*(mld1(l)**2))))-0.5+lg1(l)
      n=n+m
      n1=n1+m
      enddo

c ******************** calculate p-value ****************************************
      call sort(t1,m1)
      x4=t1(1)
      h=1
      do while(x4.lt.t)
        h=h+1
        x4=t1(h)
      enddo
      h1=m1-h
      p=h1/m1

c ********************** bca percentile method ************************************
      call sort(xbb,m1)
      x5=xbb(1)
      h2=1
      do while(yb.gt.x5)
        h2=h2+1
        x5=xbb(h2)
      enddo
      h3=h2
      m2=m1
      h3=h3/m2
      a2=(1.0/3.0)*((m*k)**(-1.0/2.0))
      ww=zinprb(h3)
      alph(1)=.025
      alph(2)=.975
      alph(3)=.050
      alph(4)=.950
      si(1)=-1.96
      si(2)=1.96
      si(3)=-1.6449
```

```
      si(4)=1.6449
      do j=1,4
        zh(j)=ww+si(j)
      enddo
      do j=1,4
        ah(j)=prbln(ww+(zh(j)/(1.0-(a2*zh(j)))))
      enddo
      do j=1,4
        r(j)=m2*ah(j)
      enddo
      do j=1,4
        rr(j)=int(r(j))
        rrd(j)=rr(j)
      enddo
      do j=1,4
        cc(j)=ww+(zh(j)/(1.0-(a2*zh(j))))
      enddo
      do j=1,4
        ck(j)=zinprb(rrd(j)/m2)
      enddo
      do j=1,4
        ck1(j)=zinprb((rrd(j)+1.0)/m2)
      enddo
      do j=1,4
       tbca(j)=xbb(rr(j))+
     c ((cc(j)-ck(j))/(ck1(j)-ck(j)))*(xbb(rr(j)+1)-xbb(rr(j)))
      enddo
      return
      END

c *******GAMMADEVIATES ************************************************
```

Returns a data set of size n from a Gamma distribution with a specified scale and shape parameter using the algorithm discussed in Chapter 2, Section 5.

```
      Subroutine gammadeviates(n1,n2,alph,b,gd)
      Double Precision u1(100000),v(10000),k,alph,a,b
      Double Precision s1,s2,x1,y,c2,u2(100000),u3(100000)
      Double Precision x(100000),gd(100000)
      Integer i,j,m,m1,r2,r1,id1,id2,n1
      m=n1*n2

      id1=-2
      id2=-1
      k=int(alph)
      a=alph-k

c ***************** calculate y ***************************************************
      doj=1,50
        v(j)=ran1(id2)
      enddo
      x1=1.0
```

```
      do j=1,k
        c2=v(j)*x1
        x1=c2
      enddo
      y=-b*(log(x1))

c *********** generate u's ************************************************************
      do j=1,100000
        u1(j)=ran1(id1)
        u2(j)=ran1(id1)
        u3(j)=ran1(id1)
      enddo

c *********** routine for gamma deviates ***********************************************
      m1=0
      r1=1
      r2=1

      do while(m1.lt.m)
        s1=(u1(r1))**(1.0/a)
        s2=(u2(r1))**(1.0/(1.0-a))
      if ((s1+s2).le.1.0) then
        x(r2)=-b*(s1*((s1+s2)**-1))*log(u3(r1))
        gd(r2)=x(r2)+y
        r1=r1+1
        r2=r2+1
        m1=m1+1
      else
        r1=r1+1
      endif
      enddo
      return
      END

c *******LTSTAT ************************************************************
```

Returns all statistical values based on a lognormal distribution discussed in Chapter 2, Section 6 including the Chebychev and MLE based upperbounds for the mean.

```
      Subroutine ltstat(m,a0,lt,ltb,slt,mln,mdln,vrln,sdln,mlmu
     c    ,mlstd,ltc,mlcv,ermln,mlsdlt,mleucl,chby,hcl,hh)
      Double Precision a0(m),lt(m),x,xx,ltb,slt
      Double Precision aa,aa1,aa2,aa3,mln,mdln,vrln,hcl,hh
      Double Precision sdln,v1,v2,v3,v4,m1,mlmu,mlvlt,mlsdlt
      Double Precision mlstd,ltc,mlcv,ermln,mleucl,chby,m2
      Integer j,m
      m1=m
      m2=m-1
      x=0.0
      xx=0.0
      do j=1,m
        lt(j)=log(a0(j))
```

```
      enddo
      do j=1,m
        x1=x+lt(j)
        x=x1
      enddo
      ltb=x/m
      do j=1,m
        x2=xx+((lt(j)-ltb)**2)
        xx=x2
      enddo
      mlvlt=xx/m
      mlsdlt=SQRT(mlvlt)
      slt=SQRT(xx/(m-1))
      aa=0.5
      aa1=-1.0*(1/(2*(m1-1)))
      aa2=2.0
      aa3=(m1-2)/(m1-1)
      mln=EXP(ltb)*gmt(m,slt,aa)
      mdln=EXP(ltb)*gmt(m,slt,aa1)
      v1=gmt(m,slt,aa2)
      v2=gmt(m,slt,aa3)
      vrln=EXP(2.0*ltb)*(v1-v2)
      v3=(gmt(m,slt,aa))**2
      v4=gmt(m,slt,aa3)
      ermln=SQRT(EXP(2.0*ltb)*(v3-v4))
      sdln=SQRT(vrln)
      mlmu=EXP(ltb+(.5*((slt**2))))
      mlstd=SQRT(EXP((2*ltb)+(slt**2))*(EXP((slt**2))-1.0))
      ltc=slt/ltb
      mlcv=mlstd/mlmu
      mleucl=EXP(ltb+(1.6449*(slt)))
      chby=mln+(4.47*ermln)
      hcl=EXP(ltb+(0.5*(slt**2))+((slt*hh)/sqrt(m2)))
      return
      END
```

c *******LOGNORMAL **********

Returns a data set of size n from a lognormal distribution with a specified scale and shape parameters. This accomplished by first generating a variable from a standard normal distribution then scaling it with a specified mean and standard deviation to produce a deviate from a $N(\mu,\sigma)$ distribution and finally exponentiating to produce a lognormal deviate.

```
      Subroutine lognormal(n2,nd,lnsd,lnxb)
      Double Precision nd(10000),normdev,lnsd,lnxb,ndd(10000)
      Double Precision x,x1,xbar,sd,n1
      Integer i,j,id,n2
      id=-5
      n1=n2
```

```
      do j=1,n2
        ndd(j)=normdev(id)
        nd(j)=EXP((lnsd*ndd(j))+lnxb)
      enddo
      return
      END

c *******PRLNORM ************************************************************
      Subroutine prlnorm(m,nmd,pln)
      Double Precision nmd(m),pln(m),c
      Integer m
      c=0.0
      do j=1,m
        c=nmd(j)
        pln(j)=prbln(c)
      enddo
      return
      END

c *******CONFIDENCEBOUND ****************************************************
      Subroutine confidencebound(k,m,xb,h,v,yp,u)
      Double Precision v,p,c1,c2,c3,y1,y2,y3,yp,u,h(12),m,k,xb
       v=2*m*k
       p=2.0
       c1=4*(sqrt(2/v))
       c2=8/v
       c3=(16.0*sqrt(p))/(sqrt(v**3))
       y1=v-(sqrt(2*v)*h(1))+(4*h(2))+(c1*(3*h(3)+2*h(4)))
       y2=c2*(6*h(5)+3*h(6)+2*h(7))
       y3=c3*(30*h(8)+9*h(9)+12*h(10)+6*h(11)+4*h(12))
       yp=y1-y2-y3
       u=(v*xb)/yp
      return
      END

c ******NEWTON*****************************************************************
```

Returns the values maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ using the Newton-Rhapson technique and all related numerical algorithms presented in Chapter 2, Section 1.

```
      Subroutine newton(q,n,xb,b,a0,m)
      Double Precision x,x1,w,wbar,a,a0,m,psi,dpsi,b,mu,xb,q(n),w1
      Integer k,l,n
      x1=0.0
      w1=0.0
      do k=1,n
       x=x1+q(k)
       x1=x
      enddo
      xb=x1/n
      do k=1,n
```

```
      w=w1+log(q(k))
      w1=w
     enddo
     wbar=w1/n
     m=log(xb)- wbar
     a0=1/(2*m)
     do i=1,100
        psi=log(a0)-(1+(1-(.1-1/(21*a0))/(a0*a0))/(6*a0))/(2*a0)
        dpsi= (1+(1+(1-((.2-1/(7*a0))/(a0*a0)))/(3*a0))/(2*a0))/a0
        a=a0-((log(a0) - psi - m)/(1/a0 - dpsi))
        a0=a
     enddo
     if(a0.ge.1.0)then
       a0=(a0*(1-(3.0/n)))+(2.0/(3.0*n))
     endif
     b=xb/a0
     return
     END
```

c *******SORT **********

Returns a data set in ascending order using Shell's method. This is a variant of a straight sort in that the data are first sorted in groups of two and then those groups in groups of two... and then a final sort as a whole group. In the final grouping each data value should be near it's final place.

```
     Subroutine sort(r,n)
     Double Precision  r(n),v
     Integer inc,i,j,n
     inc=1
1    inc=3*inc+1
     if(inc.le.n) goto 1
2    continue
       inc=inc/3
       do i=inc+1,n
         v=r(i)
         j=i
3         if(r(j-inc).gt.v) then
            r(j)=r(j-inc)
            j=j-inc
            if(j.le.inc)goto 4
          goto 3
          endif
4         r(j)=v
       enddo
     if(inc.gt.1) goto 2
     return
     End
```

c *******SCALE **********

Returns a scaled value for a gamma deviate, i.e. dividing by the shape parameter.

```
      Subroutine scale(r,b,n,rr)
      Double Precision r(n),rr(n),b
      do j=1,n
        rr(j)=(r(j))/b
      enddo
      return
      END

c *******GSER ********************************************************************

      Subroutine gser(a,x,m,gamser,gln,pz)
      Integer ITMAX
      Real EPS,ap,del,sum,gammln
      Double Precision a,gamser,x(m),gln,pz(m)
      Parameter (ITMAX=100,EPS=3.e-7)
      Integer n,m
      gln=gammln(a)
      do j=1,m
      if(x(j).le.0.)then
        if(x(j).lt.0.)pause'x<0 in gser'
        gamser=0.
        return
      endif
      ap=a
      sum=1./a
      del=sum
      do n=1,ITMAX
        ap=ap+1.
        del=del*(x(j))/ap
        sum=sum+del
        if(abs(del).lt.abs(sum)*EPS)goto 1
      enddo
      pause 'a too large, ITMAX too small in gser'
1     gamser=sum*exp(-x(j)+a*log(x(j))-gln)
      pz(j)=gamser
      enddo
      return
      END

c *******KSTAT ********************************************************************
```

Calculates the Kolmogrov-Smirnov test statistics based on the theory of Chapter 2, Section 4, to be used in determining whether the Gamma distribution with the fitted parameters provides an adequate fit.

```
      Subroutine kstat(m,p,ks)
      Double Precision p(m),k(m),k2(m),ks,s(m),s2(m),ml
      Integer m,j,i
      ml=m
      ks=0.0
```

```
      do i=1,m
        s(i)=i/m1
        s2(i)=(i-1)/m1
      enddo
      do j=1,m
        k(j)=ABS(s(j)-p(j))
        k2(j)=ABS(p(j)-s2(j))
      enddo
      call sort(k,m)
      call sort(k2,m)
      if (k(m).ge.k2(m))then
        ks=k(m)
      else
        ks=k2(m)
      endif
      return
      END

c ***************************************************************************
```

VITA

Graduate College
University of Nevada, Las Vegas

Ross J. Iaci

Local Address:
29 Rue De Parc
Henderson, Nevada 89014

Home Address:
6612 Reynard Drive
Springfield, Virginia 22152

Degrees:
Bachelor of Science, Mathematical Sciences, 1994
University of North Carolina, Chapel Hill

Special Honors and Awards:
National Collegiate Mathematics Award for excellence in mathematics

Thesis Title: The Gamma Distribution as an Alternative to the Log Normal Distribution in Environmental Applications

Thesis Examination Committee:
Chairperson, Dr. Ashok K. Singh
Committee Member, Dr. Malwane Ananda
Committee Member, Dr. Rohan Dalpatadu
Graduate Faculty Representative, Dr. Laxmi Gewali