Use of log normal transformation in environmental statistics

Sally L. Stewart

University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
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. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

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.
USE OF LOG NORMAL TRANSFORMATIONS

IN

ENVIRONMENTAL STATISTICS

by

SALLY L. STEWART

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Mathematics

Department of Mathematics
University of Nevada, Las Vegas

December, 1994
The Thesis of Sally L. Stewart for the degree of Master of Science in Mathematics is approved.

Chairperson, Ashok K. Singh, Ph. D.

Examining Committee Member, Malwane Ananda, Ph. D.

Examining Committee Member, Rohan Dalpatadu, Ph. D.

Graduate Faculty Representative, Sidkazem Taghva, Ph. D.

Dean of the Graduate College, Cheryl L. Bowles, Ed. D.

University of Nevada, Las Vegas
December, 1994
ABSTRACT

The log normal transformation is commonly used in the analysis of environmental data. The sample histogram of observed contaminant concentrations from a Superfund site typically appears to be log normal and the concentration data is log-transformed so that the classical statistical methods based on normal distribution can be used. USEPA guidance documents on statistical evaluation of attainment of cleanup standards for soils suggest using the log normal transformation in case the contaminant concentration data appears to be log normal. There are two basic problems with using a transformation in data analysis:

i) interpretation of results, and

ii) in transforming a formula based on the assumption of normality of the data so that it can be applied to transformed data.

The present thesis will address the second problem associated with the log transformation. In addition, the performance of some of the common normal-theory based procedures applied on original concentration data when the data distribution is in fact log normal will be investigated. Real Superfund site characterization data and simulated data will be used to provide examples.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>1. INTRODUCTION TO THE LOG NORMAL DISTRIBUTION</td>
<td></td>
</tr>
<tr>
<td>HISTORY</td>
<td>1</td>
</tr>
<tr>
<td>DEFINITION</td>
<td>4</td>
</tr>
<tr>
<td>MOTIVATION FOR USE</td>
<td>6</td>
</tr>
<tr>
<td>2. ESTIMATION OF THE MEAN OF A LOG NORMAL DISTRIBUTION</td>
<td></td>
</tr>
<tr>
<td>METHOD 1: POINT ESTIMATION</td>
<td>10</td>
</tr>
<tr>
<td>METHOD 2: CONFIDENCE INTERVAL ESTIMATION</td>
<td>11</td>
</tr>
<tr>
<td>NORMAL THEORY</td>
<td>12</td>
</tr>
<tr>
<td>SICHEL’S METHOD</td>
<td>12</td>
</tr>
<tr>
<td>3. ONE PROBLEM WITH THE USE OF LOG NORMAL TRANSFORMATION</td>
<td>13</td>
</tr>
<tr>
<td>4. SIMULATION EXPERIMENT</td>
<td>18</td>
</tr>
<tr>
<td>DESCRIPTION - PART 1</td>
<td>18</td>
</tr>
<tr>
<td>DESCRIPTION - PART 2</td>
<td>23</td>
</tr>
<tr>
<td>RESULTS OF SIMULATIONS EXPERIMENT</td>
<td>27</td>
</tr>
<tr>
<td>5. CORRECT METHOD OF APPLYING SAMPLE SIZE FORMULA</td>
<td>31</td>
</tr>
<tr>
<td>APPENDIX I</td>
<td></td>
</tr>
<tr>
<td>PROOFS OF LOG NORMAL FORMULAS</td>
<td>33</td>
</tr>
<tr>
<td>APPENDIX II</td>
<td></td>
</tr>
<tr>
<td>TABLE 1: SICHEL’S T-ESTIMATOR OF THE MEAN $\psi(\beta^{2}, n)$</td>
<td>35</td>
</tr>
<tr>
<td>TABLE 2: SICHEL’S T-ESTIMATOR OF LOWER 5% CI</td>
<td>36</td>
</tr>
<tr>
<td>TABLE 3: SICHEL’S T-ESTIMATOR OF UPPER 5% CI</td>
<td>37</td>
</tr>
<tr>
<td>TABLE 4: RESULTS OF SIMULATION EXPERIMENT</td>
<td>38</td>
</tr>
<tr>
<td>TABLE 5: RESULTS OF SIMULATION EXPERIMENT</td>
<td>40</td>
</tr>
<tr>
<td>APPENDIX III</td>
<td></td>
</tr>
<tr>
<td>FORTRAN PROGRAM FOR SIMULATION EXPERIMENT</td>
<td>42</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>49</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Kapteyn's Analogue Machine for Generating a Skew Frequency</td>
<td>3</td>
</tr>
<tr>
<td>2. Frequency Curves for $N(x</td>
<td>\mu=0, \sigma^2 = 0.5)$ and $\Lambda(x</td>
</tr>
<tr>
<td>3. Frequency Curves for $\Lambda(x</td>
<td>\mu = 0, \sigma^2 = 0.5)$, $\Lambda(x</td>
</tr>
<tr>
<td>4(a) Example 1: Histogram of Superfund Site 1 (raw data)</td>
<td>8</td>
</tr>
<tr>
<td>4(b) Example 1: Histogram of Superfund Site 1 (In of raw data)</td>
<td>8</td>
</tr>
<tr>
<td>5(a) Example 2: Histogram of Superfund Site 2 (raw data)</td>
<td>9</td>
</tr>
<tr>
<td>5(b) Example 2: Histogram of Superfund Site 2 (In of raw data)</td>
<td>9</td>
</tr>
<tr>
<td>6. Kolmogorov-Smirnov Test for Normality for Area A (raw data)</td>
<td>15</td>
</tr>
<tr>
<td>7(a) Kolmogorov-Smirnov Test for Normality for Area B (raw data)</td>
<td>16</td>
</tr>
<tr>
<td>7(b) Kolmogorov-Smirnov Test for Normality for Area B (In of raw data)</td>
<td>16</td>
</tr>
<tr>
<td>8. Graph of Mean Differences in Original Variables vs.</td>
<td>17</td>
</tr>
<tr>
<td>Transformed Variables ($\mu_0=1$)</td>
<td></td>
</tr>
<tr>
<td>9. Graph of Mean Differences in Original Variables vs.</td>
<td>17</td>
</tr>
<tr>
<td>Transformed Variables ($\mu_0=5$)</td>
<td></td>
</tr>
<tr>
<td>10(a) Kolmogorov-Smirnov Test for Normality for Example 1 (raw data)</td>
<td>29</td>
</tr>
<tr>
<td>10(b) Kolmogorov-Smirnov Test for Normality for Example 1 (In of raw data)</td>
<td>29</td>
</tr>
<tr>
<td>11(a) Kolmogorov-Smirnov Test for Normality for Example 2 (raw data)</td>
<td>30</td>
</tr>
<tr>
<td>11(b) Kolmogorov-Smirnov Test for Normality for Example 2 (In of raw data)</td>
<td>30</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Ashok K. Singh, for suggesting this topic, and supporting me throughout my studies. I am grateful for the opportunities which have come my way related to my area of study.

I would also like to thank Mr. Ken Brown of United States Environmental Protection Agency, for his support and help in using real-life examples from past EPA work.

I am thankful for the EBSCoR Women in Science Grant which funded my last year of study. It offered me additional time to devote to research and learning. In addition, it funded my trip to the Conference of Environmetrics in Burlington, Canada where I presented my thesis. I especially want to thank Ellen Jacobson, Director of EBSCoR, for her guidance and advice.

Lastly, I would like to thank John for all his support, encouragement, and patience throughout the year.
CHAPTER 1

INTRODUCTION TO THE LOG NORMAL DISTRIBUTION

HISTORY

The theory of log normal distribution appears to have been first introduced by D. McAlister in his memoir presented to the Royal Society of London in 1879 [14] in which he gave expressions for the mean, median, mode and the second moment of the distribution. The memoir was presented by Francis Galton, who originally suggested the study. In his opening remarks [7], Galton expressed the view that in certain cases the geometric mean is to be preferred to the arithmetic mean as a measure of location. His assumption lies at the basis of the well-known law of 'Frequency of Error' which he believes to be incorrectly applied to many social phenomena. In 1903, the next advance was made by J. C. Kapteyn [10] in which he established clearer genesis of the distribution and described a machine for generating samples from a log normal population similar to Galton for normal populations.

Kapteyn's theory on the genesis of the log normal distribution is based on the law of proportionate effect which states that a change in the variate at any step of the process is a random proportion of a function $\phi(X_{j-1})$ of the value $X_{j-1}$ already attained [10]. In other words, suppose that the variate is initially $X_0$ and that after the jth step in the process it is $X_j$; the final value is $X_n$. Then the general case suggested by Kapteyn is $X_j - X_{j-1} = e_j \phi(X_{j-1})$. However, the special case $\phi(X) = X$ (the change of the variate is a random proportion of the momentary value of the variate) of proportionate effect would reduce to $X_j - X_{j-1} = e_j X_{j-1}$. The connection between this law and the additive form of the central limit theorem is shown in the proof of Theorem 2, Appendix I. Thus, Kapteyn's machine (see Figure 1) is based on the generating model:

$$X_j - X_{j-1} = e_j X_{j-1} \ (j = 1, \ldots, n),$$

where $e_j$ is specified by:
\[P(e_j = a) = 1/2 \text{ and } P(e_j = -a) = 1/2, \text{ for all } j, \text{ and } a \text{ is a positive constant.}\]

The machine consists of nine rows of wedges encased in a wood and glass frame 104 cm high. The width of the wedges are proportional to the distance of the vertex of the wedge from the left-hand side of the frame. i.e., if \(X_{j-1}\) is the distance of a vertex from the left-hand side of the frame, the width of the wedge is \(2aX_{j-1}\).

Sand is poured into a funnel directly above the center wedge in the top row. When arriving at the point \(X_{j-1}\), the sand is divided into two equal parts, displaced either to \(X_{j-1}(1 + a)\) or \(X_{j-1}(1 - a)\). The sand arrives in the receptacles at the bottom of the machine, forming a skewed histogram.

M. J. Van Uven joined J. C. Kapteyn in 1916 to further develop the distribution in which estimation using quantiles was added [11]. Shortly after, the distribution received great criticism from K. Pearson who based his objections on general mistrust of the technique of transformations. Interest in the distribution died down until about 1930 when papers published by Clark [4], Hemmingsen [8], and Bliss [2] indicated that log normal distributions were effective in normalizing distributions in biological studies. With the invention of high-speed computing, sophisticated methods of analysis were developed in order to create tables of characteristics of the log normal distribution. The United States Environmental Protection Agency has currently developed two packages, SCOUT and GEO-EAS [16], which include the log transformation of data. In this thesis, the SCOUT package is used to perform the Kolmogorov-Smirnov test for normality, and the GEO-EAS package to compute sample statistics. The characteristics of the two-parameter log normal distribution are defined below.
Figure 1. Kapteyn's Analogue Machine for Generating a Skew Frequency
DEFINITION

Consider a positive variate \( Y \) \((0 < y < \infty)\) such that \( X = \ln Y \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \). Then we say that \( Y \) is log normally distributed with parameters \( \mu, \sigma \), and denote it by \( Y \sim \Lambda(y \mid \mu, \sigma^2) \) and correspondingly, \( X \sim N(x \mid \mu, \sigma^2) \). It is important to note that the distribution of \( X \) is completely specified by the two parameters \( \mu, \sigma \). However, \( Y \) cannot assume zero values since the transformation \( X = \ln Y \) is not defined for \( Y = 0 \). Figure 2 gives a comparison of the frequency curves of the \( N(x \mid \mu=0, \sigma^2=0.5) \) and \( \Lambda(y \mid \mu=0, \sigma^2=0.5) \), showing the positions of the mean, median and mode for the \( \Lambda(y \mid \mu=0, \sigma^2=0.5) \) distribution. Since \( X \) and \( Y \) are connected by the relationship \( X = \ln Y \), the distribution functions are related. Thus, from the properties of the moment generating function of the normal distribution, we can derive the following formulas (see Appendix I) [1]:

- Mean of \( Y \) = \( E(Y) = e^{\mu + 0.5\sigma^2} \)
- Median of \( Y \) = \( e^\mu \)
- Mode of \( Y \) = \( e^{\mu - \sigma^2} \)
- Variance of \( Y \) = \( \text{VAR}(Y) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) = (e^{\mu + 0.5\sigma^2})^2 \eta^2 \), where \( \eta^2 = (e^{\sigma^2} - 1) \)
- Skewness of \( Y \) = \( \eta^3 + 3\eta \)
- Coefficient of Kurtosis of \( Y \) = \( \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 \)

It is clear from the above formulas that the distribution is positively skewed and that the greater the value of \( \sigma^2 \), the greater the skewness. Also, the distribution has positive kurtosis which increases as \( \sigma^2 \) increases. Figure 3 shows the frequency curves for \( \Lambda(x \mid \mu = 0, \sigma^2 = 0.5) \), \( \Lambda(x \mid \mu = 0, \sigma^2 = 0.1) \), and \( \Lambda(x \mid \mu = 0, \sigma^2 = 2) \) from which the flexibility of the distribution may be obtained. An additional remark is that the two-parameter log normal distribution has several properties which are immediate consequences of those for the normal distribution. In particular, it is important to note that \( E[\ln(Y)] \neq \ln(E[Y]) \).
Figure 2. Frequency Curves for $N(x | \mu = 0, \sigma^2 = 0.5)$ and $\Lambda(x | \mu = 0, \sigma^2 = 0.5)$

Figure 3. Frequency Curves $\Lambda(x | \mu = 0, \sigma^2 = 0.5)$, $\Lambda(x | \mu = 0, \sigma^2 = 0.1)$, and $\Lambda(x | \mu = 0, \sigma^2 = 2)$
MOTIVATION FOR USE

The log normal distribution can be adequately described in natural occurrences of observed distributions in several fields of study, such as Economics, Biology, and Small-particle Statistics. Documented examples of application of log normal theory in these fields follow, with emphasis on Small-particle Statistics.

In the field of Economics, distributions of personal income have attracted the greatest attention. The choice of a particular form of the distribution is governed by the statistical description of the model and the criterion specified. Champemowne [3] developed a model which depended on the subdivision of income into discrete ranges. Contrarily, Lorenz [12] developed a model based on the concept of the concentration of incomes. The evidence studied by the authors suggested that the distribution of income is in fact log normal. Moreover, the more homogeneous the group of income recipients is, the more likely the distribution is log normal.

In the field of Biology, Cramer [5] discusses the growth of an organism subject to a number of small independent impulses acting in an ordered sequence. The law of proportionate effect applies if the influence of each impulse is proportionate to the momentary size of the organism. Thus, the final size of the organism will tend to be log normally distributed (as proved prior by Kapteyn).

The log normal distribution is well-established in Small-particle statistics. Many contamination data sets are highly skewed with as much as hundred-fold increase in size from the smallest to the largest. In addition, researchers are often interested in related particle measurements such as diameters, volumes and weights. Extensive research by Matheron [13] showed that the geochemical process of solution and concentration tends to produce log normal distributions for grades in mining applications.
The following two examples demonstrate that contaminant concentration data is typically highly skewed.

**EXAMPLE 1:** Samples of PCB collected from Superfund Site 1

Statistical analysis was performed on the data samples collected from Site 1 using The Environmental Protection Agency's GEO-EAS package. As we can see from Figure 4(a), the coefficient of skewness is 1.730 which implies that the data is highly skewed. However, by using the transformation $X = \ln Y$, and performing the analysis on the transformed data, we can see from Figure 4(b) the coefficient of skewness is now -0.522 which means the data is just slightly skewed.

**EXAMPLE 2:** Samples of Chrysene collected from Superfund Site 2

Using the same statistical package (GEO-EAS), Figure 5(a) analysis shows the coefficient of skewness for the original data is 1.848 which implies the data is highly skewed. However, once the data is transformed, we can see in Figure 5(b) that the data is just slightly skewed since the coefficient of skewness is -0.328.
Figure 4(a). Example 1: Histogram of Superfund Site 1 (raw data)

Figure 4(b). Example 1: Histogram of Superfund Site 1 (ln of raw data)
Figure 5(a). Example 2: Histogram of Superfund Site 2 (raw data)

Figure 5(b). Example 2: Histogram of Superfund Site 2 (ln of raw data)
CHAPTER 2

ESTIMATION OF THE MEAN OF A LOG NORMAL DISTRIBUTION

METHOD 1: POINT ESTIMATION

Generally, a random variable $X$ has a probability density function of known form which depends on an unknown parameter $\theta$, $\theta \in \Omega$. Therefore, we have a family of distributions for each value of $\theta$, $\theta \in \Omega$. We denote this family as $\mathcal{F} = \{f(x; \theta) : \theta \in \Omega\}$. Our goal is to select one member of this family as being the probability density function of $X$. In other words, we want a point estimate of $\theta$. This estimate of $\theta$ is denoted by $\hat{\theta}$. Following are some of the desirable properties which a good estimator must possess.

i) the estimator should be unbiased, i.e., $E(\hat{\theta}) = \theta$.

ii) the variance of the estimator should be minimized.

Suppose we are given $y_1, y_2, \ldots, y_n$ independent random samples from $\mathcal{N}(\mu, \sigma^2)$. The sample mean, $\bar{y} = 1/n \sum_{i=1}^{n} y_i$, can be used as an estimate of $\mu$. However, it is known to be an inefficient estimator of $\mu$ since it is usually affected by a few large values.

For data that is highly skewed, the arithmetic mean of assays tends to overestimate the mean of the distribution due to the presence of erratic high values. Sichel [15] theoretically derived a better way to estimate the mean by the following relationship: $\mu = e^{\bar{x} + \frac{3}{2} s^2}$, where $\bar{x}$ is the mean of the logarithms and $s$ is their standard deviation. However, this relationship is between parameter values and not their estimates. Sichel developed the Sichel "t-estimator" to overcome this estimation problem. It is defined as follows:
\[ t = e^y \psi \left( \frac{s^2}{2}, n \right), \text{ where } \psi(u,n) = 1 + \frac{n-1}{n} u + \frac{(n-1)^3}{n^2(n+1)2.1} u^2 + \ldots \text{ and} \]

\[ s^2 = \text{MLE of } \sigma^2. \]

Tables of \( \psi(\beta^2, n) \) (Table 1, Appendix II) are given in Sichel [15]. Sichel's estimator is similar in form to the minimum variance unbiased estimator derived by Finney [6].

To demonstrate the use of Sichel's Table, suppose we have a sample of size 10 randomly drawn from a log normal distribution. The logarithmic mean, \( \alpha = 5.298 \) and the variance, \( \beta^2 = 0.8 \). Then the mean of \( \bar{X} \) is given as:

\[ \bar{X} = e^y \psi(\beta^2, n), \text{ where } \beta^2 = 0.8, n = 10 \]

\[ = e^{5.298(1.472)} \quad \text{(from Table 1, Appendix II)} \]

\[ = 294.3 \]

**METHOD 2: CONFIDENCE INTERVAL ESTIMATION**

The confidence interval utilizes the information in the sample to arrive at two numbers that are intended to enclose the parameter, \( \theta \), of interest. Ideally, we would like the interval to have two properties: (i) the interval should contain the target parameter, \( \theta \); and (ii) the interval should be relatively narrow. The probability that a confidence interval will enclose \( \theta \) is called the confidence coefficient, denoted by \( 1 - \alpha \). This confidence coefficient gives the fraction of the time, in repeated sampling, that the interval constructed will contain the target parameter \( \theta \).

Suppose that \( \theta_L \) and \( \theta_U \) denote the lower and upper confidence limits, respectively, for a parameter \( \theta \). If \( P(\theta_L < \theta < \theta_U) = 1 - \alpha \), then the probability, \( (1 - \alpha) \), is the confidence coefficient. The random interval, \((\theta_L, \theta_U)\), is called a two-sided confidence interval.

For example, let \( \alpha = 0.05 \). If we performed repeated sampling, say 100 times, then 95% of the
time, our confidence interval would contain our target parameter $\theta$. In other words, in the long run, about 95 out of the 100 confidence intervals constructed would contain $\theta$.

**NORMAL THEORY**

Since we rarely know the form of the population frequency distribution, we make the assumption that the samples have been randomly selected from a normal population, where $\mu$ and $\sigma^2$ are unknown. Also, we know that $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$ has a t-distribution with $(n-1)$ degrees of freedom.

Then we can form a confidence interval for $\mu$. The resulting confidence interval is: $\bar{Y} \pm t_{5\alpha} \frac{S}{\sqrt{n}}$.

**SICHEL'S METHOD**

Sichel calculated multiplying factors to compute a central 90% confidence interval for the mean of the log normal distribution. Table 2 and 3 in Appendix II is the lower and upper 5% limits of error of the $t$-estimator, respectively.

As an illustration of the use of the tables to compute a 90% confidence interval, suppose we have a sample of size 10 randomly drawn from a log normal distribution. The logarithmic mean, $\bar{Y} = 5.298$ and the variance, $\hat{\sigma}^2 = 0.8$. Then the 90% confidence interval for the mean is given as:

$C_L = e^{u_t_L} (\hat{\sigma}^2, n), \text{where } \hat{\sigma}^2 = 0.8, n = 10$

$= e^{5.298(0.93)}$ (from Table 2, Appendix II)

$= 186.0$

$C_U = e^{u_t_U} (\hat{\sigma}^2, n), \text{where } \hat{\sigma}^2 = 0.8, n = 10$

$= e^{5.298(3.55)}$ (from Table 3, Appendix II)

$= 709.8$
CHAPTER 3

ONE PROBLEM WITH THE USE OF LOG NORMAL DISTRIBUTIONS

The United States Environmental Protection Agency (USEPA) Guidance document [17] provides detailed information on statistical methods applicable to the problem of deciding, on the basis of a random sample collected from the site, whether the site meets the cleanup standards or not. In particular, this guidance document provides the following formula for determining the number of samples required to obtain a specified confidence level:

\[
    n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(C_s - \mu_1)^2},
\]

where \( C_s \) = cleanup standard for the site
\( \sigma^2 \) = variance (estimated)
\( \mu_1 \) = mean under the alternative hypothesis (< \( C_s \))
\( z_{1-\alpha} \) = upper 100(1-\( \alpha \))% point of standard normal distribution
\( z_{1-\beta} \) = upper 100(1-\( \beta \))% point of standard normal distribution

An evaluation of an EPA Superfund site was requested to determine the number of boring locations necessary to characterize chemical concentrations of sediments in the study area. The above-referenced USEPA evaluation method utilizing the expected variability based on historic data was used.

Preliminary evaluation of the results from the analysis of the USACE samples indicated that two distinct 2,3,7,8-TCDD concentration distributions were present. Thus, the site was divided into two areas: A and B, respectively. The following historical data was used in the computations:

<table>
<thead>
<tr>
<th>AREA A (ppt)</th>
<th>49</th>
<th>180</th>
<th>290</th>
<th>220</th>
<th>230</th>
<th>110</th>
<th>57</th>
<th>120</th>
<th>94</th>
<th>120</th>
<th>210</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA B (ppt)</td>
<td>760</td>
<td>480</td>
<td>1500</td>
<td>68</td>
<td>380</td>
<td>6300</td>
<td>20</td>
<td>820</td>
<td>960</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The evaluation of the data was performed by first computing summary statistics and secondly, evaluating the underlying distribution of the data. The summary statistics are as follows:

<table>
<thead>
<tr>
<th>AREA</th>
<th>y</th>
<th>s</th>
<th>MIN</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>153</td>
<td>78</td>
<td>49</td>
<td>290</td>
</tr>
<tr>
<td>B</td>
<td>1734</td>
<td>2242</td>
<td>20</td>
<td>6300</td>
</tr>
</tbody>
</table>

The Kolmogorov-Smirnov test was used for Area A (see Figure 6) and did not reject normality for Distribution A. Thus, Area A’s underlying distribution is indicated to be normal. On the other hand, the Kolmogorov-Smirnov test rejected normality for Distribution B (see Figure 7(a)). However, by converting the samples of Area B to log-scale, the Kolmogorov-Smirnov test did not reject normality for the log-scale data (see Figure 7(b)). Thus, the underlying distribution for Area B is indicated to be log normal.

Using the above-mentioned formula for the number of boring locations needed and given the following criteria:

\[ \text{AREA A: } \alpha = 0.05 \quad \beta = 0.20 \quad C_s \cdot \mu_1 = 80 \text{ ppt} \]

\[ \text{AREA B: } \alpha = 0.05 \quad \beta = 0.20 \quad C_s \cdot \mu_1 = 100 \text{ ppt} \]

the contractor suggested that a total of eight samples per mile would be adequate to characterize 2378-TCDD sediment concentrations for Area A. Additionally, seventeen samples per mile would be adequate for Area B. However, this estimation appeared to be incorrect because the formula for the number of samples depends on the standard deviation, which is considerably higher for Area B. Consequently, the number of samples for Area B was calculated using the standard deviation of the real scale values of the samples and it was determined that 3054 samples would be adequate.

This huge discrepancy in the two sample sizes forced us to look at the usage of the sample-size formula on log-transformed data. In this example, the contractor replaced the error limit of 100 ppt by ln(100). The problem with this approach is that the difference of ln(100) on a log-scale does not
translate to a difference of 100 in the means of the concentrations on the real scale. This is shown clearly by the following two graphs. In Figure 8, the X-axis represents the difference of mean of log-transformed variables ($\mu_1 - \mu_0$) when $\mu_0 = 1$. The Y-axis represents the difference in means of the original variables. If we look at a change of log(100) from $\mu_0$ in the X-direction, this corresponds to a change of 100 in the Y-direction. However, in Figure 9, a change of log(100) in the X-direction corresponds to a change of nearly 8000 in the Y-direction.

Figure 6. Kolmogorov-Smirnov Test for Normality for Area A (raw data)
Figure 7(a). Kolmogorov-Smirnov test for normality for Area B (raw data)

Figure 7(b). Kolmogorov-Smirnov Test for Normality for Area B (ln of raw data)
Figure 8. Graph of Mean Differences in Original Variables vs. Transformed Variables
($\mu_0 = 1$)

Figure 9. Graph of Mean Differences in Original Variables vs. Transformed Variables
($\mu_0 = 5$)
CHAPTER 4

SIMULATION EXPERIMENT

As we have shown, contaminant concentration data is typically highly skewed. The statistician normally will compute the Kolmogorov-Smirnov test for normality to determine if the sample data fit a normal distribution. If not, the sample data will be log transformed. The Kolmogorov-Smirnov test will then be computed on the transformed data to determine if it fits a normal distribution. If so, "normal" theory-based formulas will be applied for the analysis of the sample data.

We know that for a small number of sample data, it is very difficult to determine what type of distribution the data may have. On the other hand, for a large number of samples, we can apply the Central Limit Theorem. The question that arises, then, is: "For a small number of sample data, is it necessary to log transform the sample data and apply Sichel's theory to determine an estimate for the mean, or rather, use Normal theory estimates on the real-scale sample points?" As discussed prior, there is misinterpretation of the use of this log transformed sample data. The goal of the simulation experiment was to compare these two methods of confidence interval estimation for the mean in hopes to answer the question proposed.

A FORTRAN program was written and included in Appendix III for the simulation experiment in which Monte Carlo simulation was used to generate sample data. The simulation experiment is fully described below.

DESCRIPTION - PART 1

The first step in the simulation experiment was to generate a random sample of size \( n \) from a log normal distribution with parameters \( \mu, \sigma \) [denoted \( \Lambda(y \mid \mu, \sigma^2) \)]. This was accomplished using IMSL STAT/LIBRARY [9] subroutine RNLLNL. Once our samples were generated, a 90% confidence
interval was computed using "normal" theory. Then each sample was log transformed. A 90% confidence interval was then computed on the transformed samples using "Sichel's" theory. Since Sichel's t-tables were limited, we determined whether or not the confidence interval could be computed due to missing table values. This is indicated on the output as "Number of Misses on Table". Once the confidence interval for each method was computed, a determination of whether the true mean was contained in each interval was noted. This is indicated as "Number of Hits for Normal Theory" and "Number of Hits for Log Normal Theory", for each method respectively. Finally, the average interval length and its standard deviation was computed.

In Examples 1, 2 and 3 that follow, Step 1 of the simulation experiment is demonstrated. Note that the average length of the intervals for each theory is the actual interval length which has a standard deviation value as zero since we have only computed one interval for each method.

In Example 1, we have requested 10 samples from a $\Lambda(y \mid \mu = 1.1, \sigma^2 = .16)$. First, the output shows what we have requested and lists the generated data points. The 90% confidence interval is computed. The data points are then log transformed and listed. The 90% confidence interval under Sichel's theory is computed. The results of Example 1 show that the true mean is contained in each interval.

Similarly, in Example 2, we have requested 10 samples from a $\Lambda(y \mid \mu = 1.1, \sigma^2 = .25)$. The results are the same as in Example 1.

Example 3 was generated from a $\Lambda(y \mid \mu = 1.1, \sigma^2 = .04)$. The results for this example show that an interval under Sichel's theory was not computed due to missing table value. Therefore, the interval defaults to $[0,0]$ and the indicator for number of misses on table is now 1. The true mean is contained in the interval computed under normal theory.
EXAMPLE 1

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND A STANDARD DEVIATION: 0.40000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.254375

THE GENERATED LOG NORMAL DATA POINTS ARE:

3.157331  2.747720  5.457298  5.348978  1.618213
3.098199  3.111766  1.753176  5.046471  4.516624

THE 90% CONFIDENCE INTERVAL BASED ON NORMAL THEORY IS:
[2.803816, 4.367339]

THE LN OF THE GENERATED DATA POINTS ARE:

1.149727  1.010772  1.696954  1.676905  0.4813222
1.130821  1.135190  0.5614293  1.618689  1.507765

THE 90% CONFIDENCE INTERVAL BASED ON SICHEL'S THEORY IS:
[2.945927, 5.329149]

THE NUMBER OF HITS FOR NORMAL THEORY IS: 1
AVERAGE LENGTH: 1.5635  S.D: 0.0000

THE NUMBER OF HITS FOR LN THEORY IS: 1
AVERAGE LENGTH: 2.3832  S.D: 0.0000
NUMBER OF MISSES ON TABLE IS: 0

THIS IS THE END OF THE RUN
EXAMPLE 2

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND A STANDARD DEVIATION: 0.50000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.404166

THE GENERATED LOG NORMAL DATA POINTS ARE:

4.882174  2.535231  1.617043  2.480418  1.570380
11.17985  12.99840  6.926432  1.563053  3.284885

THE 90% CONFIDENCE INTERVAL BASED ON NORMAL THEORY IS:
[ 2.613038, 7.194535]

THE LN OF THE GENERATED DATA POINTS ARE:

1.585591  0.9302849  0.4805991  0.9084271  0.4513175
2.414113  2.564826  1.935345  0.4466408  1.189332

THE 90% CONFIDENCE INTERVAL BASED ON SICHEL’S THEORY IS:
[ 3.271626, 10.06934]

THE NUMBER OF HITS FOR NORMAL THEORY IS: 1
AVERAGE LENGTH: 4.5815  S.D: 0.0000

THE NUMBER OF HITS FOR LN THEORY IS: 1
AVERAGE LENGTH: 6.7977  S.D: 0.0000

NUMBER OF MISSES ON TABLE IS: 0

THIS IS THE END OF THE RUN
EXAMPLE 3

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND STANDARD DEVIATION: 0.20000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.064854

THE GENERATED LOG NORMAL DATA POINTS ARE:

| 3.526819 | 4.589726 | 2.952541 | 2.681173 | 2.466662 |
| 2.262289 | 3.188480 | 3.437556 | 3.396833 | 3.197326 |

THE 90% CONFIDENCE INTERVAL BASED ON NORMAL THEORY IS:
[ 2.798257, 3.530624]

THE LN OF THE GENERATED DATA POINTS ARE:

| 1.260396 | 1.523820 | 1.082666 | 0.9862543 | 0.9028659 |
| 0.8163772 | 1.159544 | 1.234761 | 1.222844 | 1.162315 |

THE 90% CONFIDENCE INTERVAL BASED ON SICHEL'S THEORY IS:
[ 0.0000, 0.0000]

THE NUMBER OF HITS FOR NORMAL THEORY IS: 1
AVERAGE LENGTH: 0.7214 S.D: 0.0000

THE NUMBER OF HITS FOR LN THEORY IS: 0
AVERAGE LENGTH: 0.0000 S.D: 0.0000
NUMBER OF MISSES ON TABLE IS: 1

THIS IS THE END OF THE RUN
DESCRIPTION - PART 2

Recall that a specified confidence of 90% means that for repeated samplings, about 90% of the estimated intervals should contain the true mean. Consequently, for the second part of the simulation experiment, Step 1 was repeatedly run the desired number of times. The output has been modified to reflect only the computed confidence intervals for each simulation run without listing the actual simulated data points for each run. Therefore, the indicators for the "Number of Hits for Normal Theory" and "Number of Hits for Log Normal Theory", for each method respectively, reflect the total for all the simulation runs along with the average interval length and its standard deviation.

Examples 4, 5, and 6, correspond to Examples 1, 2 and 3, respectively, with the number of simulations increased to ten. Note that the first interval in Examples 4, 5 and 6, are exactly the same as in their corresponding example.

The results of Example 4 indicates that the true mean was contained in all 10 of the intervals computed for both methods.

In Example 5, we see that 9 out of 10 of the intervals contained the true mean for Sichel's theory, whereas all 10 did for normal theory.

Example 6 shows that under Sichel's theory, 9 intervals were not computed due to missing table values, but the remaining one did contain the true mean. Under normal theory, all 10 intervals did contain the true mean.
EXAMPLE 4

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND STANDARD DEVIATION: 0.40000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.254375

THE 90% CONFIDENCE INTERVAL

<table>
<thead>
<tr>
<th>NORMAL THEORY</th>
<th>SICHEL'S THEORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ 2.8038, 4.3673]</td>
<td>[ 2.9459, 5.3291]</td>
</tr>
<tr>
<td>[ 2.3935, 3.6491]</td>
<td>[ 2.5408, 3.8395]</td>
</tr>
<tr>
<td>[ 2.1225, 3.5190]</td>
<td>[ 2.3449, 3.5434]</td>
</tr>
<tr>
<td>[ 2.5602, 3.6278]</td>
<td>[ 2.6635, 4.0249]</td>
</tr>
<tr>
<td>[ 2.7610, 4.7020]</td>
<td>[ 3.0642, 5.5432]</td>
</tr>
<tr>
<td>[ 2.2100, 3.2571]</td>
<td>[ 2.3485, 3.5488]</td>
</tr>
<tr>
<td>[ 2.5450, 3.6758]</td>
<td>[ 2.5860, 4.6780]</td>
</tr>
<tr>
<td>[ 3.0108, 3.9093]</td>
<td>[ 3.0353, 4.5866]</td>
</tr>
<tr>
<td>[ 2.6643, 4.1545]</td>
<td>[ 2.8094, 5.0822]</td>
</tr>
<tr>
<td>[ 2.7008, 4.6110]</td>
<td>[ 2.9391, 5.3168]</td>
</tr>
</tbody>
</table>

THE NUMBER OF HITS FOR NORMAL THEORY IS: 10
AVERAGE LENGTH: 1.3701 S.D: 0.3404

THE NUMBER OF HITS FOR LN THEORY IS: 10
AVERAGE LENGTH: 1.8215 S.D: 0.5163
NUMBER OF MISSES ON TABLE IS: 0

THIS IS THE END OF THE RUN
EXAMPLE 5

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND STANDARD DEVIATION: 0.50000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.404166

THE 90% CONFIDENCE INTERVAL
FOR

<table>
<thead>
<tr>
<th>NORMAL THEORY</th>
<th>SICHEL'S THEORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6130, 7.1945</td>
<td>3.2716, 10.0693</td>
</tr>
<tr>
<td>2.5770, 3.9915</td>
<td>2.7570, 4.1661</td>
</tr>
<tr>
<td>2.7860, 4.7012</td>
<td>3.0390, 5.4975</td>
</tr>
<tr>
<td>2.1868, 4.6069</td>
<td>2.5350, 5.4975</td>
</tr>
<tr>
<td>2.2167, 3.4240</td>
<td>2.3241, 4.2042</td>
</tr>
<tr>
<td>2.8889, 4.2500</td>
<td>2.9700, 5.3728</td>
</tr>
<tr>
<td>2.9972, 5.0091</td>
<td>3.2026, 5.7935</td>
</tr>
<tr>
<td>2.6777, 4.5829</td>
<td>2.9260, 5.2931</td>
</tr>
<tr>
<td>2.8945, 5.2829</td>
<td>3.9403, 9.4302</td>
</tr>
<tr>
<td>2.8945, 5.2829</td>
<td>3.2350, 5.8520</td>
</tr>
</tbody>
</table>

THE NUMBER OF HITS FOR NORMAL THEORY IS: 10
AVERAGE LENGTH: 2.3715 S.D: 1.1535

THE NUMBER OF HITS FOR LN THEORY IS: 9
AVERAGE LENGTH: 3.0856 S.D: 1.6041
NUMBER OF MISSES ON TABLE IS: 0

THIS IS THE END OF THE RUN
EXAMPLE 6

REQUIRED INPUT: Number of samples, Number of simulation runs, Mean, Standard deviation, T-value, T-column

PROGRAM OUTPUT: The program returns the following output.

YOU HAVE REQUESTED 10 DATA POINTS
GENERATED FROM A LN SAMPLE WITH MEAN: 1.10000
AND STANDARD DEVIATION: 0.20000

THE EXACT MEAN BASED ON LOG NORMAL THEORY FOR THIS DATA IS: 3.064854

THE 90% CONFIDENCE INTERVAL
FOR
NORMAL THEORY   SICHEL'S THEORY
[ 2.8093, 3.5306]     [ 0.0000, 0.0000]
[ 2.8119, 3.3804]     [ 0.0000, 0.0000]
[ 2.6725, 3.2991]     [ 0.0000, 0.0000]
[ 2.7759, 3.3546]     [ 0.0000, 0.0000]
[ 2.8161, 3.1976]     [ 0.0000, 0.0000]
[ 2.6470, 3.3807]     [ 0.0000, 0.0000]
[ 3.0041, 3.3246]     [ 0.0000, 0.0000]
[ 2.5416, 3.1179]     [ 0.0000, 0.0000]
[ 2.9843, 3.7493]     [ 0.0000, 0.0000]
[ 2.8375, 3.7257]     [ 2.8767, 4.3470]

THE NUMBER OF HITS FOR NORMAL THEORY IS: 10
AVERAGE LENGTH: 0.6160  S.D: 0.1640

THE NUMBER OF HITS FOR LN THEORY IS: 1
AVERAGE LENGTH: 1.4703  S.D: 0.0000

NUMBER OF MISSES ON TABLE IS: 9

THIS IS THE END OF THE RUN
RESULTS OF SIMULATION EXPERIMENT

The simulation experiment was repeated 100, 1000, and 10000 times for the requested number of data samples [5, 10, 15, 20, 25, 30, 50, 100]. Table 4 in Appendix II is a summary of the simulation experiment results. The columns are divided by the number of data samples requested and subdivided by the number of simulations runs. The rows are divided by standard deviation and subdivided by the two estimation methods. Also, the number of uncalculated estimated intervals for Sickle's theory was noted. Each cell entry represents either the number of times the estimated interval contained the true mean under the appropriate theory or the number of uncalculated intervals for Sichel's theory. For example, the results state for a standard deviation of .4 and 5 data samples requested, 86 out of 100 simulation runs contained the true mean for "normal" theory, whereas 71 out of 100 simulation runs contained the true mean with 24 intervals not computed for Sichel's theory.

The results of our extensive simulation experiment indicates that the approximate "normal" theory based confidence interval compares quite well in comparison to Sichel's theory based confidence interval for the mean of the log normal distribution. Overall, the average interval length for normal theory was smaller (see Table 5, Appendix II) with a higher confidence percentage (see Table 4, Appendix II).

For small n, it is not easy to test for normality or log-normality. By computing the Kolmogorov-Smirnov test for normality on the sample points in Example 1, the results indicate that the sample points fit a normal distribution (see Figure 10(a)). In addition, Figure 10(b) shows that the log transformation of the sample points also fit a normal distribution. Likewise, the same results follow with the sample data in Example 2 (see Figure 11(a),and (b)).

Since the approximate "normal" theory based confidence interval performs quite well even for small n, we might want to just use the approximate normal theory based methods.
For large \( n \), the approximate "normal" theory based confidence interval is almost identical to Sichel's theory based confidence interval and hence, it does not pay to log-transform data to attain normality.

If log transformation must be used, then care should be taken in transforming formulas based on normal theory.

The simulation experiment was modified to generate data samples from a normal distribution with parameters \( \mu, \sigma \), for varying values of the parameters. In each case, the normal theory-based intervals were shorter in length and contained the true mean more times than that of Sichel's theory.

In conclusion, if the experimenter knows that the underlying distribution is log normal, and the parameters of interest are the natural log normal parameters \( (\mu, \sigma^2) \), then log normal theory should be used. However, if the interest is in the mean of the distribution, then it is not necessary to log transform the sample data set when estimating the true mean of the sample distribution.
Figure 10(a). Kolmogorov-Smirnov Test for Normality for Example 1 (raw data)

Figure 10(b). Kolmogorov-Smirnov Test for Normality for Example 1 (ln of raw data)
Figure 11(a). Kolmogorov-Smirnov Test for Normality for Example 2 (raw data)

Figure 11(b). Kolmogorov-Smirnov Test for Normality for Example 2 (ln of raw data)
CHAPTER 5

CORRECT METHOD OF APPLYING SAMPLE SIZE FORMULA

From Chapter 3, the formula to determine the number of samples is:

\[ n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(C_s - \mu_1)^2} \]

where \( C_s \) = cleanup standard for the site
\( \sigma^2 \) = variance (estimated)
\( \mu_1 \) = mean under the alternative hypothesis (< \( C_s \))
\( z_{1-\alpha} \) = upper 100(1-\( \alpha \))% point of standard normal distribution
\( z_{1-\beta} \) = upper 100(1-\( \beta \))% point of standard normal distribution

We determined that the error limit, \( (C_s - \mu_1) \), cannot be replaced by simply just taking the log transformation of the error limit. Recall that our test to determine whether a site is clean is given by:

\[ H_0: E(Y) = C_s \quad \text{vs} \quad H_{a1}: E(Y) = C_s - \Delta \]

We wish to detect the difference \( \Delta \) in the mean of the y-values. The difference is then:

\[ \Delta = C_s - E(Y \mid H_{a1}) = e^{\mu_1 + 0.5\sigma^2} - e^{\mu_1 + 0.5\sigma^2}, \text{where } \sigma^2 = \text{sample variance of } X = \ln(Y). \]

Next, from the clean-up standard, \( C_s = e^{\mu_1 + 0.5\sigma^2} \), we have:

\( \mu_0 = \ln(C_s) - 0.5\sigma^2; \text{and from } E(Y \mid H_{a0}) = M_1 = e^{\mu_1 + 0.5\sigma^2}, \text{we have:} \)

\( \mu_1 = \ln(M_1) - 0.5\sigma^2 \)

Thus, \( \mu_0 - \mu_1 = \ln(C_s) - 0.5\sigma^2 - [\ln(M_1) - 0.5\sigma^2] = \ln\left(\frac{C_s}{M_1}\right) \neq \ln(\Delta) \)

Therefore, the corrected formula for the number of samples for log transformed data is given by:

\[ n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\ln\left(\frac{C_s}{M_1}\right)^2} \]

where all variables are defined as above.
One point that need be clarified is that it is not enough to specify the $\Delta$ alone, the clean-up standard must be specified. If we review the requirements for the sample design given in the problem of Chapter 3, the error limit for Area B was 100 ppt. That is $C_S - \mu_1 = 100$ ppt.

Case 1: Let $C_S = 200, M_1 = 100$

then the number of samples required is given by: 
$$n = \frac{(2.487)^2 (3.447)}{\ln(200/100)^2} = 45$$

Case 2: Let $C_S = 5000, M_1 = 4900$

then the number of samples required is given by: 
$$n = \frac{(2.487)^2 (3.447)}{\ln(5000/4900)^2} = 52,237$$

The error limit in each case is 100 ppt, but the number of samples required is dramatically different.
APPENDIX I

PROOFS OF LOG NORMAL FORMULAS

Theorem 1 [1]: Let \( \{X_i\} \) be a sequence of independent, positive variates such that
\[
E(\log X_j) = \mu_j, \quad D^2(\log X_j) = \sigma_j^2, \quad \text{and} \quad E(\mid \log X_j - \mu_j |^3) = \omega_j^3 \quad \text{all exist for every } j.
\]
Then if
\[
\lim_{j \to \infty} \omega_j = \infty.
\]

then the product \( \prod_{j=1}^{n} X_j \) is asymptotically distributed as \( \Lambda(\mu_n, \sigma_n^2) \), provided
\[
\omega_n / \sigma_n \to 0, \quad \text{as } n \to \infty.
\]

Theorem 2 [1]: A variate subject to the law of proportionate effect tends, for large \( n \), to be distributed as a two-parameter \( \Lambda - \text{variate} \), provided that the sequence \( X_0, 1 + \varepsilon_1, 1 + \varepsilon_2, \ldots \), satisfies the conditions of Theorem 1.

Proof: Given: \( X_j - X_{j-1} = \varepsilon_j X_{j-1} \) the law of proportionate effect. We can rewrite this as:
\[
\frac{X_j - X_{j-1}}{X_{j-1}} = \varepsilon_j \quad \text{so that} \quad \sum_{j=1}^{n} \frac{X_j - X_{j-1}}{X_{j-1}} = \sum_{j=1}^{n} \varepsilon_j.
\]
If the effect at each step is small, we have
\[
\sum_{j=1}^{n} \frac{X_j - X_{j-1}}{X_{j-1}} = \int_{X_0}^{X_n} \frac{dX}{X} = \log(X_n) - \log(X_0)
\]
giving \( \log X_n = \log X_0 + \varepsilon_1 + \ldots + \varepsilon_n \). By the additive form of the central-limit theorem \( \log X_n \) is asymptotically normally distributed and hence \( X_n \) is asymptotically log normally distributed in a two-parameter form.
PROOF OF LOG NORMAL PROPERTIES

Since \( X \) and \( Y \) are related by \( X = \ln Y \), the distribution function of \( X \) and \( Y \) are related by:

\[ A(y) = A(\ln y), \text{ where } (y > 0) \text{ and } A(y) = 0, (y \leq 0). \]

Thus, the distribution function of \( Y \) is:

\[ dA(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\ln y - \mu)^2\right)dy, \quad (y > 0) \]

The distribution possesses moments of any order, the jth moment about the origin is denote by:

\[ \lambda_j = \int_0^\infty y^j dA(y) = \int_0^\infty e^{jt} dN(x) = e^{j\mu + 0.5j\sigma^2}. \]

Therefore, the mean = \( E(Y) \) is derived using the first moment:

\[ E(Y) = \lambda_1 = e^{\mu + 0.5\sigma^2}. \]

Similarly, we can derive the variance, coefficient of skewness, and coefficient of kurtosis.
## APPENDIX II

### TABLE 1: SICHEL'S T-ESTIMATOR OF THE MEAN $\psi(\beta^2, n)$

<table>
<thead>
<tr>
<th># of samples/var ln(x)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
<td>1.051</td>
</tr>
<tr>
<td>0.2</td>
<td>1.158</td>
<td>1.105</td>
<td>1.105</td>
<td>1.105</td>
<td>1.105</td>
<td>1.105</td>
<td>1.105</td>
<td>1.105</td>
</tr>
<tr>
<td>0.3</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
<td>1.274</td>
</tr>
<tr>
<td>0.4</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
</tr>
<tr>
<td>0.5</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
<td>1.523</td>
</tr>
<tr>
<td>0.6</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
<td>1.651</td>
</tr>
<tr>
<td>0.7</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
<td>1.798</td>
</tr>
<tr>
<td>0.8</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
<td>1.954</td>
</tr>
<tr>
<td>0.9</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
<td>2.116</td>
</tr>
<tr>
<td>1.0</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
<td>2.278</td>
</tr>
<tr>
<td>1.1</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
<td>2.448</td>
</tr>
<tr>
<td>1.2</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
<td>2.624</td>
</tr>
<tr>
<td>1.3</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
<td>2.806</td>
</tr>
<tr>
<td>1.4</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
<td>3.000</td>
</tr>
<tr>
<td>1.5</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
<td>3.197</td>
</tr>
<tr>
<td>1.6</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
<td>3.401</td>
</tr>
<tr>
<td>1.9</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
<td>4.032</td>
</tr>
<tr>
<td>2.0</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
<td>4.251</td>
</tr>
<tr>
<td>2.3</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
<td>4.946</td>
</tr>
<tr>
<td>2.4</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
<td>5.192</td>
</tr>
<tr>
<td>2.6</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
<td>5.705</td>
</tr>
<tr>
<td>2.7</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
<td>5.971</td>
</tr>
<tr>
<td>3.1</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
<td>7.093</td>
</tr>
<tr>
<td>3.2</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
<td>7.388</td>
</tr>
<tr>
<td>3.3</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
<td>7.689</td>
</tr>
<tr>
<td>3.4</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
<td>7.996</td>
</tr>
<tr>
<td>3.6</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
<td>8.628</td>
</tr>
<tr>
<td>3.7</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
<td>8.954</td>
</tr>
</tbody>
</table>
### TABLE 2: SICHEL'S T-ESTIMATOR ($t_L$) FOR LOWER 5% CONFIDENCE INTERVAL

<table>
<thead>
<tr>
<th># of samples/\text{var ln(x)}</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.83</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>0.2</td>
<td>0.79</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.3</td>
<td>0.76</td>
<td>0.89</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.4</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.5</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.6</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.7</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.8</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.9</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.0</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.1</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.2</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.4</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.74</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
</tr>
</tbody>
</table>
TABLE 3: SICHEL'S T-ESTIMATOR (t_U) FOR UPPER 5% CONFIDENCE INTERVAL

<table>
<thead>
<tr>
<th>#of samples/var ln(x)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.74</td>
<td>1.36</td>
<td>1.27</td>
<td>1.21</td>
<td>1.17</td>
<td>1.15</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>2.30</td>
<td>1.61</td>
<td>1.45</td>
<td>1.38</td>
<td>1.34</td>
<td>1.31</td>
<td>1.28</td>
<td>1.20</td>
</tr>
<tr>
<td>0.3</td>
<td>2.94</td>
<td>1.86</td>
<td>1.64</td>
<td>1.53</td>
<td>1.48</td>
<td>1.44</td>
<td>1.40</td>
<td>1.39</td>
</tr>
<tr>
<td>0.4</td>
<td>3.67</td>
<td>2.13</td>
<td>1.84</td>
<td>1.70</td>
<td>1.63</td>
<td>1.57</td>
<td>1.52</td>
<td>1.49</td>
</tr>
<tr>
<td>0.5</td>
<td>4.54</td>
<td>2.43</td>
<td>2.05</td>
<td>1.88</td>
<td>1.78</td>
<td>1.72</td>
<td>1.66</td>
<td>1.62</td>
</tr>
<tr>
<td>0.6</td>
<td>5.57</td>
<td>2.77</td>
<td>2.26</td>
<td>2.07</td>
<td>1.96</td>
<td>1.88</td>
<td>1.82</td>
<td>1.79</td>
</tr>
<tr>
<td>0.7</td>
<td>6.80</td>
<td>3.13</td>
<td>2.53</td>
<td>2.27</td>
<td>2.12</td>
<td>2.03</td>
<td>1.95</td>
<td>1.88</td>
</tr>
<tr>
<td>0.8</td>
<td>8.30</td>
<td>3.55</td>
<td>2.81</td>
<td>2.50</td>
<td>2.31</td>
<td>2.20</td>
<td>2.11</td>
<td>2.04</td>
</tr>
<tr>
<td>0.9</td>
<td>10.00</td>
<td>4.00</td>
<td>3.11</td>
<td>2.74</td>
<td>2.54</td>
<td>2.40</td>
<td>2.31</td>
<td>2.24</td>
</tr>
<tr>
<td>1.0</td>
<td>12.10</td>
<td>4.52</td>
<td>3.44</td>
<td>3.00</td>
<td>2.77</td>
<td>2.61</td>
<td>2.50</td>
<td>2.43</td>
</tr>
<tr>
<td>1.1</td>
<td>14.60</td>
<td>5.02</td>
<td>3.80</td>
<td>3.38</td>
<td>3.08</td>
<td>2.84</td>
<td>2.70</td>
<td>2.63</td>
</tr>
<tr>
<td>1.2</td>
<td>17.60</td>
<td>5.60</td>
<td>4.20</td>
<td>3.69</td>
<td>3.34</td>
<td>3.05</td>
<td>2.90</td>
<td>2.84</td>
</tr>
<tr>
<td>1.3</td>
<td>21.20</td>
<td>6.40</td>
<td>4.60</td>
<td>4.03</td>
<td>3.50</td>
<td>3.20</td>
<td>3.07</td>
<td>3.01</td>
</tr>
<tr>
<td>1.4</td>
<td>25.60</td>
<td>7.30</td>
<td>5.10</td>
<td>4.30</td>
<td>3.80</td>
<td>3.60</td>
<td>3.50</td>
<td>3.43</td>
</tr>
<tr>
<td>1.5</td>
<td>30.60</td>
<td>8.20</td>
<td>5.60</td>
<td>4.70</td>
<td>4.10</td>
<td>3.90</td>
<td>3.80</td>
<td>3.73</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>90</td>
<td>875</td>
<td>8642</td>
<td>82</td>
<td>678</td>
<td>8821</td>
<td>83</td>
</tr>
<tr>
<td>m = 2.0</td>
<td>LN</td>
<td>20</td>
<td>178</td>
<td>1745</td>
<td>14</td>
<td>198</td>
<td>1850</td>
<td>11</td>
</tr>
<tr>
<td>m = 3.0</td>
<td>LN</td>
<td>80</td>
<td>820</td>
<td>8234</td>
<td>85</td>
<td>801</td>
<td>8098</td>
<td>88</td>
</tr>
<tr>
<td>m = 4.0</td>
<td>LN</td>
<td>89</td>
<td>837</td>
<td>8578</td>
<td>86</td>
<td>878</td>
<td>8763</td>
<td>83</td>
</tr>
<tr>
<td>m = 5.0</td>
<td>LN</td>
<td>67</td>
<td>552</td>
<td>5788</td>
<td>79</td>
<td>723</td>
<td>7427</td>
<td>80</td>
</tr>
<tr>
<td>m = 6.0</td>
<td>LN</td>
<td>29</td>
<td>423</td>
<td>3986</td>
<td>16</td>
<td>227</td>
<td>2119</td>
<td>14</td>
</tr>
<tr>
<td>m = 7.0</td>
<td>LN</td>
<td>86</td>
<td>845</td>
<td>8514</td>
<td>96</td>
<td>876</td>
<td>8644</td>
<td>87</td>
</tr>
<tr>
<td>m = 8.0</td>
<td>LN</td>
<td>71</td>
<td>749</td>
<td>7707</td>
<td>97</td>
<td>873</td>
<td>8749</td>
<td>89</td>
</tr>
<tr>
<td>m = 9.0</td>
<td>LN</td>
<td>24</td>
<td>203</td>
<td>1816</td>
<td>0</td>
<td>44</td>
<td>386</td>
<td>0</td>
</tr>
<tr>
<td>m = 10.0</td>
<td>LN</td>
<td>83</td>
<td>840</td>
<td>8278</td>
<td>80</td>
<td>867</td>
<td>8665</td>
<td>87</td>
</tr>
<tr>
<td>m = 11.0</td>
<td>LN</td>
<td>83</td>
<td>861</td>
<td>8375</td>
<td>81</td>
<td>892</td>
<td>8926</td>
<td>86</td>
</tr>
<tr>
<td>m = 12.0</td>
<td>LN</td>
<td>12</td>
<td>81</td>
<td>958</td>
<td>2</td>
<td>11</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>m = 13.0</td>
<td>LN</td>
<td>80</td>
<td>818</td>
<td>8204</td>
<td>93</td>
<td>829</td>
<td>8459</td>
<td>87</td>
</tr>
<tr>
<td>m = 14.0</td>
<td>LN</td>
<td>85</td>
<td>859</td>
<td>8733</td>
<td>93</td>
<td>887</td>
<td>8979</td>
<td>92</td>
</tr>
<tr>
<td>m = 15.0</td>
<td>LN</td>
<td>2</td>
<td>49</td>
<td>476</td>
<td>0</td>
<td>3</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>m = 16.0</td>
<td>LN</td>
<td>75</td>
<td>813</td>
<td>8062</td>
<td>84</td>
<td>850</td>
<td>8328</td>
<td>88</td>
</tr>
<tr>
<td>m = 17.0</td>
<td>LN</td>
<td>90</td>
<td>886</td>
<td>6904</td>
<td>89</td>
<td>905</td>
<td>8976</td>
<td>89</td>
</tr>
<tr>
<td>m = 18.0</td>
<td>LN</td>
<td>4</td>
<td>26</td>
<td>308</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>m = 19.0</td>
<td>LN</td>
<td>79</td>
<td>785</td>
<td>7800</td>
<td>81</td>
<td>808</td>
<td>8101</td>
<td>72</td>
</tr>
<tr>
<td>m = 20.0</td>
<td>LN</td>
<td>90</td>
<td>894</td>
<td>6792</td>
<td>86</td>
<td>881</td>
<td>8943</td>
<td>82</td>
</tr>
<tr>
<td>m = 21.0</td>
<td>LN</td>
<td>3</td>
<td>25</td>
<td>348</td>
<td>0</td>
<td>6</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>m = 22.0</td>
<td>LN</td>
<td>76</td>
<td>765</td>
<td>7528</td>
<td>85</td>
<td>817</td>
<td>7986</td>
<td>77</td>
</tr>
<tr>
<td>m = 23.0</td>
<td>LN</td>
<td>84</td>
<td>869</td>
<td>8474</td>
<td>85</td>
<td>888</td>
<td>8806</td>
<td>88</td>
</tr>
<tr>
<td>m = 24.0</td>
<td>LN</td>
<td>4</td>
<td>53</td>
<td>624</td>
<td>3</td>
<td>16</td>
<td>233</td>
<td>0</td>
</tr>
<tr>
<td>m = 25.0</td>
<td>LN</td>
<td>71</td>
<td>725</td>
<td>7396</td>
<td>78</td>
<td>761</td>
<td>7745</td>
<td>75</td>
</tr>
<tr>
<td>m = 26.0</td>
<td>LN</td>
<td>84</td>
<td>808</td>
<td>8054</td>
<td>82</td>
<td>836</td>
<td>8327</td>
<td>83</td>
</tr>
<tr>
<td>m = 27.0</td>
<td>LN</td>
<td>7</td>
<td>100</td>
<td>1056</td>
<td>4</td>
<td>79</td>
<td>748</td>
<td>7</td>
</tr>
<tr>
<td>m = 28.0</td>
<td>LN</td>
<td>55</td>
<td>453</td>
<td>4629</td>
<td>50</td>
<td>518</td>
<td>5129</td>
<td>55</td>
</tr>
<tr>
<td>m = 29.0</td>
<td>LN</td>
<td>16</td>
<td>148</td>
<td>1572</td>
<td>1</td>
<td>30</td>
<td>292</td>
<td>1</td>
</tr>
<tr>
<td>m = 30.0</td>
<td>LN</td>
<td>73</td>
<td>764</td>
<td>7440</td>
<td>91</td>
<td>924</td>
<td>9160</td>
<td>97</td>
</tr>
</tbody>
</table>

<p>| TABLE 4: RESULTS OF SIMULATION EXPERIMENT |
| (DATA POINTS = 15, 16, 15, 20) |</p>
<table>
<thead>
<tr>
<th>m</th>
<th>sd</th>
<th>NORMA L</th>
<th>25</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100 / 1,000 / 10,000</td>
<td>100 / 1,000 / 10,000</td>
<td>100 / 1,000 / 10,000</td>
<td>100 / 1,000 / 10,000</td>
<td></td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>85</td>
<td>883</td>
<td>8833</td>
<td>91</td>
<td>900</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>13</td>
<td>142</td>
<td>1489</td>
<td>9</td>
<td>133</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>87</td>
<td>848</td>
<td>8458</td>
<td>90</td>
<td>864</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>90</td>
<td>906</td>
<td>8871</td>
<td>89</td>
<td>880</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>90</td>
<td>871</td>
<td>8808</td>
<td>94</td>
<td>890</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>90</td>
<td>862</td>
<td>8996</td>
<td>92</td>
<td>868</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>87</td>
<td>868</td>
<td>8761</td>
<td>86</td>
<td>891</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>89</td>
<td>901</td>
<td>9018</td>
<td>87</td>
<td>893</td>
</tr>
<tr>
<td>m = 1.2</td>
<td>NORMAL</td>
<td>91</td>
<td>892</td>
<td>8748</td>
<td>86</td>
<td>861</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>95</td>
<td>914</td>
<td>9090</td>
<td>90</td>
<td>914</td>
</tr>
<tr>
<td>m = 1.2</td>
<td>NORMAL</td>
<td>84</td>
<td>856</td>
<td>8632</td>
<td>92</td>
<td>836</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>90</td>
<td>898</td>
<td>9085</td>
<td>95</td>
<td>882</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>84</td>
<td>845</td>
<td>8491</td>
<td>87</td>
<td>856</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>89</td>
<td>902</td>
<td>9015</td>
<td>90</td>
<td>898</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>89</td>
<td>824</td>
<td>8433</td>
<td>75</td>
<td>815</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>90</td>
<td>896</td>
<td>9004</td>
<td>89</td>
<td>872</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>85</td>
<td>820</td>
<td>8247</td>
<td>83</td>
<td>808</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>89</td>
<td>870</td>
<td>8740</td>
<td>86</td>
<td>888</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>61</td>
<td>565</td>
<td>339</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>2</td>
<td>34</td>
<td>5927</td>
<td>58</td>
<td>643</td>
</tr>
<tr>
<td>m = 1.1</td>
<td>NORMAL</td>
<td>99</td>
<td>999</td>
<td>9956</td>
<td>100</td>
<td>999</td>
</tr>
</tbody>
</table>

**Table 4: Results of Simulation Experiment**

(Data Points = 25, 30, 50, 100)
TABLE 5: RESULTS OF SIMULATION EXPERIMENT

<table>
<thead>
<tr>
<th>m</th>
<th>sd</th>
<th>NORMAL</th>
<th>LN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.07</td>
<td>0.9770</td>
<td>0.9810</td>
</tr>
<tr>
<td>2.0</td>
<td>1.02</td>
<td>1.0238</td>
<td>1.0234</td>
</tr>
<tr>
<td>3.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>4.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>5.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>6.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>7.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>8.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>9.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
<tr>
<td>10.0</td>
<td>3.04</td>
<td>3.0313</td>
<td>3.0313</td>
</tr>
</tbody>
</table>

**INTERVAL LENGTHS (DATA POINTS = 15,16,15,20)**
<table>
<thead>
<tr>
<th>m</th>
<th>25</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>0.7983</td>
<td>0.7890</td>
<td>0.7673</td>
<td>0.7459</td>
</tr>
<tr>
<td>2.2</td>
<td>0.7868</td>
<td>0.7697</td>
<td>0.7433</td>
<td>0.7111</td>
</tr>
<tr>
<td>2.3</td>
<td>0.7475</td>
<td>0.7328</td>
<td>0.7052</td>
<td>0.6773</td>
</tr>
<tr>
<td>2.4</td>
<td>0.7205</td>
<td>0.7053</td>
<td>0.6733</td>
<td>0.6420</td>
</tr>
<tr>
<td>3.1</td>
<td>0.6341</td>
<td>0.6095</td>
<td>0.5743</td>
<td>0.5391</td>
</tr>
<tr>
<td>4.1</td>
<td>0.5056</td>
<td>0.4790</td>
<td>0.4451</td>
<td>0.4116</td>
</tr>
<tr>
<td>4.2</td>
<td>0.5051</td>
<td>0.4780</td>
<td>0.4442</td>
<td>0.4106</td>
</tr>
<tr>
<td>4.3</td>
<td>0.5048</td>
<td>0.4779</td>
<td>0.4440</td>
<td>0.4105</td>
</tr>
<tr>
<td>4.4</td>
<td>0.5046</td>
<td>0.4777</td>
<td>0.4439</td>
<td>0.4104</td>
</tr>
<tr>
<td>4.5</td>
<td>0.5045</td>
<td>0.4776</td>
<td>0.4438</td>
<td>0.4104</td>
</tr>
<tr>
<td>4.6</td>
<td>0.5044</td>
<td>0.4774</td>
<td>0.4437</td>
<td>0.4104</td>
</tr>
<tr>
<td>4.7</td>
<td>0.5043</td>
<td>0.4773</td>
<td>0.4436</td>
<td>0.4104</td>
</tr>
<tr>
<td>4.8</td>
<td>0.5042</td>
<td>0.4772</td>
<td>0.4435</td>
<td>0.4104</td>
</tr>
<tr>
<td>4.9</td>
<td>0.5041</td>
<td>0.4771</td>
<td>0.4434</td>
<td>0.4104</td>
</tr>
<tr>
<td>5.1</td>
<td>0.4796</td>
<td>0.4451</td>
<td>0.4116</td>
<td>0.3781</td>
</tr>
<tr>
<td>6.1</td>
<td>0.4128</td>
<td>0.3781</td>
<td>0.3436</td>
<td>0.3081</td>
</tr>
</tbody>
</table>

**TABLE 5: RESULTS OF SIMULATION EXPERIMENT**
INTERVAL LENGTHS (DATA POINTS = [25, 30, 50, 100])
APPENDIX III

FORTRAN PROGRAM FOR SIMULATION EXPERIMENT

************************************************************************************
* PROGRAM NAME: RANDLNSM.F
*
* THIS PROGRAM GENERATES THE DESIRED NUMBER OF DATA SAMPLES FROM
* A LOGNORMAL DISTRIBUTION WITH THE SPECIFIED MEAN AND STANDARD
* DEVIATION. THE EXACT MEAN IS COMPUTED (EXP(XM + 1/2 S^2). THE
* MEAN OF THE SAMPLE IS THEN ESTIMATED WITH A SPECIFIED
* CONFIDENCE INTERVAL USING "NORMAL" THEORY.
* THE SAMPLE POINTS ARE THEN LOG TRANSFORMED. THE MEAN OF THE
* TRANSFORMED POINTS IS ESTIMATED USING SICHEL'S THEORY.
* THE TOTAL NUMBER OF TIMES THE EXACT MEAN IS CONTAINED IN EACH
* EACH INTERVAL IS RECORDED.
*
************************************************************************************

PROGRAM MAINLN

** INITIALIZE VARIABLES

REAL XM, S, R(100), NRSQ, SLNM, XMEAN, XS, TVAL, XLNM, SS
REAL SSUM, XSD, LNR(100), LNXS, LNXMEAN, LNNS, LNSSUM
REAL LNXXVAR, LOW5TAB(16,8), HI5TAB(16,8), CLNLO, CLNUP
REAL TLO, TUP, INTLN(10000), INTLLN(10000)
REAL NOSIMSQ, LNS, LLNS, XNLEN, XLNLEN, SS1, SS2
REAL SS1SUM, SS2SUM, XSDNL, XSDLNL, MISSSQ
INTEGER NR, INSEED, NOUT, CI, NOSIM, TCOL, TROW
INTEGER NHITS, LNHITS, MISS
EXTERNAL RNLNL, RNSET, UMACH

** LOAD TABLES - lower 5% then upper 5%

DATA LOW5TAB/ .83,.79,.76,.74,.74,.74,.73,.74,.74,.74,
/ .75,.76,.77,.77,0,
/ .90,.89,.88,.89,.90,.91,.93,.94,.96,.97,
/ .99,1.01,1.03,1.05,0,
/ .93,.92,.94,.95,.98,99,1.01,1.03,1.05,1.08,
/ 1.10,1.13,1.15,1.18,0,
/ .94,.95,.96,.97,.99,1.01,1.04,1.06,1.09,1.11,1.14,
/ 1.17,1.20,1.23,1.27,0,
/
**SET INPUT DATA**

*  
  INSEED = 52334  
  NR = 10  
  NOSIM = 10  
  XM = 1.1  
  S = 0.2  
  CI = 90  
  TVAL = 1.833  
  TCOL = 2  
  NHITS = 0  
  LNHITS = 0  
  MISSES = 0  
  NRSQ = NR  
  NRSQ = SQRT (NRSQ)  
  NOSIMSQ = NOSIM  
  NOSIMSQ = SQRT (NOSIMSQ)  

**COMPUTE AND DISPLAY ENTERED INPUT AND EXACT MEAN**  
XLNM = EXP(XM + 0.5*S*S)  
PRINT 99990, NR  
99990 FORMAT (' ', 'YOU HAVE REQUESTED ', I3, ' DATA POINTS')
PRINT *, 'GENERATED FROM A LN SAMPLE WITH MEAN: ', XM
PRINT *, ' AND STANDARD DEVIATION: ', S
PRINT *, ' THE EXACT MEAN BASED ON LOGNORMAL THEORY FOR 
/ THIS DATA IS: ', XLNM
PRINT *
**
* USE THIS PRINT IF YOU WANT TO PRINT BOTH CI RESULTS TOGETHER
PRINT *, ' THE 90% CONFIDENCE INTERVAL' 
PRINT *, ' FOR' 
PRINT *, ' NORMAL THEORY        LN THEORY'
* PRINT *, '

DO 5 K = 1, NOSIM
** RESET COUNTERS
XS = 0
SSUM = 0
LNXS = 0
LNSSUM = 0
LNS = 0
LLNS = 0
SS1SUM = 0
SS2SUM = 0

**** GENERATE DATA POINTS
CALL UMACH (2,NOUT)
CALL RNSET (INSEED)
CALL RNLNL (NR, XM, S, R)

* USING THIS PRINT IF YOU WANT TO PRINT GENERATED DATA POINTS
* PRINT *, ' THE GENERATED LOGNORMAL DATA POINTS ARE:
* PRINT *, ' 
* PRINT *, (R(I), I=1,NR)

*** GENERATE A SPECIFIED CONFIDENCE INTERVAL USING "NORMAL" 
THEORY

** COMPUTE MEAN AND STANDARD DEVIATION
DO 15 I = 1, NR
XS = XS + R(I)
15 CONTINUE
XMEAN = XS / NR

DO 20 I = 1, NR
SS = (R(I) - XMEAN)**2
SSUM = SSM + SS
20 CONTINUE
XSD = SQRT (SSUM) / NRSQ

** COMPUTE AND PRINT THE CONFIDENCE INTERVAL

       CUP = XMEAN + TVAL*(XSD / NRSQ)
       CLO = XMEAN - TVAL*(XSD / NRSQ)
       INTLN(K) = CUP - CLO

USE THIS PRINT IF YOU WANT TO DISPLAY CI OF NORMAL W/GENERATED POINTS
* PRINT 99995, CI
*99995 FORMAT (' ', THE ',12,'% CONFIDENCE INTERVAL YOU REQUESTED IS:')
* PRINT *, [' ', CLO, ',', ', CUP, ']'

*** COMPUTE CI BASES ON LOGNORMAL THEORY

* TAKE LOG OF DATA

     DO 30 I = 1, NR
           LNR(I) = LOG(R(I))
     30 CONTINUE

USE THIS PRINT IF YOU WANT TO DISPLAY TRANSFORMED DATA POINTS
* PRINT *, 'THE LN OF THE GENERATED DATA POINTS ARE: '
* PRINT *, ' '
* PRINT *, (LNR(I), I = 1, NR)

** COMPUTE MEAN AND S.D. FOR LN DATA

     DO 40 I = 1, NR
           LNXS = LNXS + LNR(I)
     40 CONTINUE
     LNXMEAN = LNXS / NR

     DO 50 I = 1, NR
           LNSS = (LNR(I) - LNXMEAN)**2
           LNSSUM = LNSSUM + LNSS
     50 CONTINUE
     LNXVAR = LNSSUM / NR

*** COMPUTE CI USING LN THEORY

*** LOOK UP T FACTORS FOR UPPER/LOWER CI IN EACH TABLE

IF (LNXVAR .LE. 0.05) THEN
* PRINT *, 'LNXVAR SMALLER THAN TABLE'
  MISSES = MISSES + 1
  TROW = 16
ELSEIF (LNXVAR .LE. 0.15) THEN
  TROW = 1
ELSEIF (LNXVAR .LE. 0.25) THEN
  TROW = 2
ELSEIF (LNXVAR .LE. 0.35) THEN
  TROW = 3
ELSEIF (LNXVAR .LE. 0.45) THEN
  TROW = 4
ELSEIF (LNXVAR .LE. 0.55) THEN
  TROW = 5
ELSEIF (LNXVAR .LE. 0.65) THEN
  TROW = 6
ELSEIF (LNXVAR .LE. 0.75) THEN
  TROW = 7
ELSEIF (LNXVAR .LE. 0.85) THEN
  TROW = 8
ELSEIF (LNXVAR .LE. 0.95) THEN
  TROW = 9
ELSEIF (LNXVAR .LE. 1.05) THEN
  TROW = 10
ELSEIF (LNXVAR .LE. 1.15) THEN
  TROW = 11
ELSEIF (LNXVAR .LE. 1.25) THEN
  TROW = 12
ELSEIF (LNXVAR .LE. 1.35) THEN
  TROW = 13
ELSEIF (LNXVAR .LE. 1.45) THEN
  TROW = 14
ELSEIF (LNXVAR .LE. 1.55) THEN
  TROW = 15
ELSE
* PRINT *, 'LNXVAR LARGER THAN TABLE'
  MISSES = MISSES + 1
  TROW = 16
ENDIF
TUP = HI5TAB(TROW,TCOL)
TLO = LOW5TAB(TROW,TCOL)
CLNUP = EXP(LNXMEAN) * TUP
CLNLO = EXP(LNXMEAN) * TLO
INTLLN(K) = CLNUP - CLNLO

* USE THIS PRINT IF YOU WANT TO PRINT CI FOR BOTH THEORY TOGETHER
* PRINT *,
* PRINT 99997, CLO,CUP,CLNLO,CLNUP
*99997 FORMAT( ',[,]F8.4,[,]F8.4,[,]F8.4,[,]F8.4,[,]F8.4,[,]F8.4,[,]F8.4,[,]F8.4,)
* PRINT 99996, CI
*99996 FORMAT (', THE ',I2,'% CONFIDENCE INTERVAL BASED ON LN
* / THEORY IS:')
* PRINT *, ' ', CLNLO, ',', CLNUP, '

** COUNT NUMBER OF HITS FOR EXACT MEAN

   IF (XLM .GE. CLO) THEN
      IF (XLM .LE. CUP) THEN
         NHITS = NHITS + 1
      ENDIF
   ENDF

   IF (XLM .GE. CLNLO) THEN
      IF (XLM .LE. CLNUP) THEN
         LNHITS = LNHITS + 1
      ENDIF
   ENDF

   INSEED = INSEED + 10112
** REPEAT SIMULATION RUN THE DESIRED NUMBER OF TIMES
   5 CONTINUE

*** STEP 2 -- COMPUTE FINAL RESULTS

** COMPUTE MEAN AND STANDARD DEV OF INTERVAL LENGTHS

   DO 60 I = 1, NOSIM
      LNS = LNS + INTLN(I)
      LLNS = LLNS + INTLLN(I)
   60 CONTINUE

*** MODIFY NUMBER OF VALID INTERVALS FOR LOGNORMAL THEORY

   MISSSQ = NOSIM - MISSES
   MISSSQ = SQRT (MISSSQ)

***********

   XNLEN = LNS / NOSIM
   IF (NOSIM .GT. MISSES) THEN
      XLNLEN = LLNS / (NOSIM - MISSES)
   DO 70 I = 1, NOSIM
      SS1 = (INTLN(I) - XNLEN)**2
      SS1SUM = SS1SUM + SS1
      IF (INTLLN(I) .GT. 0) THEN
         SS2 = (INTLLN(I) - XLNLEN)**2
         SS2SUM = SS2SUM + SS2
      ENDIF
   70 CONTINUE
CON TINUE

\[ XSDNL = \frac{\text{SQRT}(SS1SUM)}{\text{NOSIMS}} \]
\[ XSDLNL = \frac{\text{SQRT}(SS2SUM)}{\text{MISSSQ}} \]
ELSE
\[ XSDLNL = 0 \]
ENDIF

** PRINT OUT FINAL RESULTS

**********

PRINT *, 'THE NUMBER OF HITS FOR NORMAL THEORY IS: ', NHITS
PRINT 99998, XNLEN, XSDNL

99998 FORMAT(','AVERAGE LENGTH: ',F8.4,10X,'S.D: ',F8.4)
PRINT *, 'THE NUMBER OF HITS FOR LN THEORY IS: ', LNHITS
PRINT 99998, XLNLEN, XSDLNL

PRINT *, 'NUMBER OF MISSES ON TABLE IS: ', MISSES
PRINT *, 'THIS IS THE END OF THE RUN'
END

****************************************************************************************************
REFERENCES


