1-1-2007

On the use of lognormal distribution for environmental data analysis

Devarshi Pant
University of Nevada, Las Vegas

NOTE TO USERS

The CD is not included in this original manuscript.

This reproduction is the best copy available.

UMI®

ON THE USE OF LOGNORMAL DISTRIBUTION FOR ENVIRONMENTAL DATA ANALYSIS

by

Devarshi Pant

Bachelor of Engineering
Shivaji University, India
2001

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2005

UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

18th Nov, 2005

The Thesis prepared by

Devarshi Pant

Entitled

On the Use of Lognormal Distribution for Environmental Data Analysis

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

M.S. MATHEMATICAL SCIENCE

Ashoon C. Sule
Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

On The Use Of Lognormal Distribution For Environmental Data Analysis

by

Devarshi Pant

Dr. A.K. Singh, Examination Committee Chair
Professor of Statistics
University of Nevada, Las Vegas

Contaminant concentration data from Superfund sites is quite often positively skewed, and the log-normal theory based statistical procedures are typically used for such data. Recent work in the environmental statistics literature, however, has shown that the use of log-normal theory based formulas, such as the H-statistic confidence interval, is problematic. The performance of the H – UCL in the presence of non – detects in the sample is investigated via simulated examples. When comparing mean contaminant concentration at a site with that of the background, the 2-sample t-test on log-transformed data is commonly used. A part of this thesis deals with investigation of power of the t-test on log-transformed data by using Monte Carlo simulation.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. A K Singh, for always being there whenever I needed his guidance. I would like to thank Dr. Anita Singh for her support and help. I wish to thank the members of my Graduate Committee (Dr. Rohan Dalpatadu, Dr. Dennis Murphy, and Dr. Laxmi Gewali) for their helpful suggestions.

I also would like to give my special thanks to Dr. Dieudonne Phanord for making this possible.

The research support from the Center of Applied Mathematics and Statistics (CAMS) of the Department of Mathematical Sciences at UNLV was helpful in completing this thesis.

I also would like to thank my parents and my uncle in India for making this happen.

CHAPTER 1

A BRIEF HISTORY OF THE NORMAL AND LOG-NORMAL DISTRIBUTIONS

This thesis is primarily concerned with the usage of log-normal distribution in environmental applications. Since the normal distribution and the log-normal distribution are closely related, a brief history of these two probability models is included in the thesis.

Normal distribution:

Abraham De Moivre, an 18th century probabilist and a consultant to gamblers was often called upon to make lengthy computations involving binomial probabilities. De Moivre observed that when the number of events (coin flips) increased, the shape of the binomial distribution approached a very smooth curve. Binomial distributions for 2, 4, and 16 tosses of a fair coin are shown in Figures 1-3.

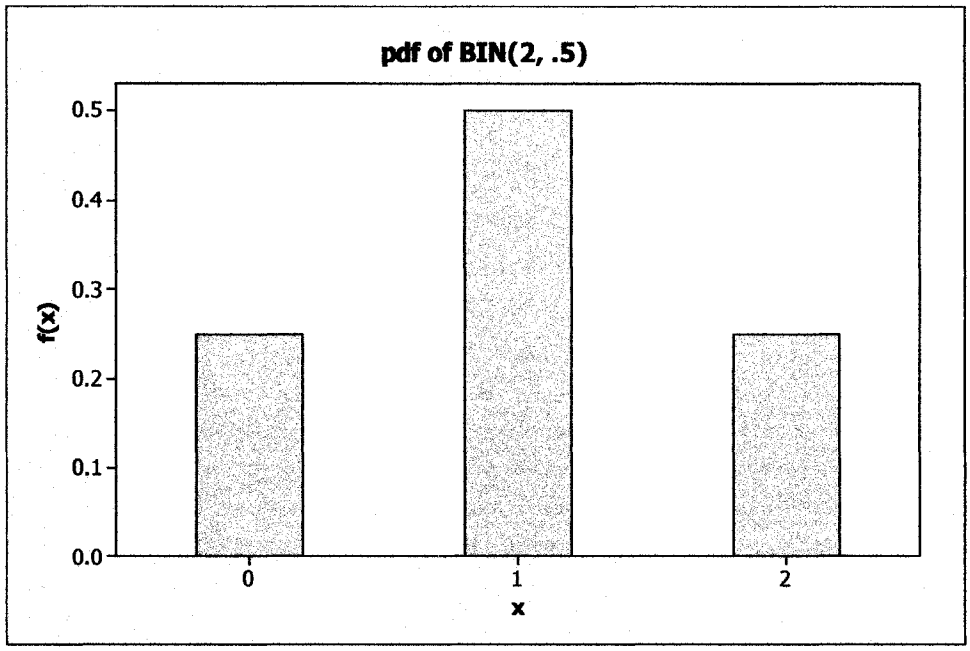

Figure 1: Graph of the binomial distribution BIN (n, p) for n = 2, p = .5

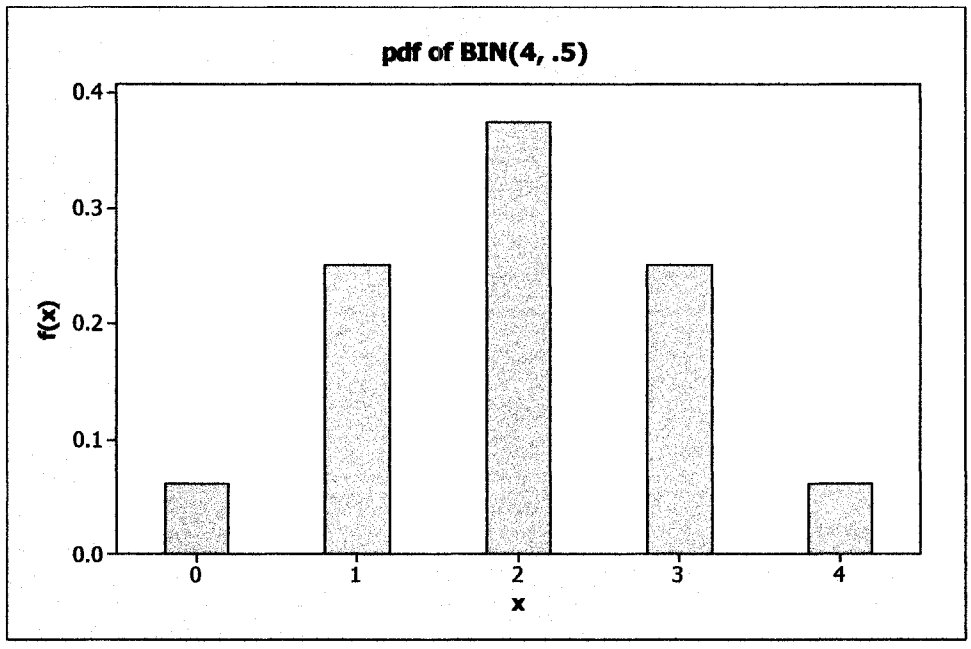

Figure 2: Graph of the binomial distribution BIN (n, p) for n = 4, p = .5

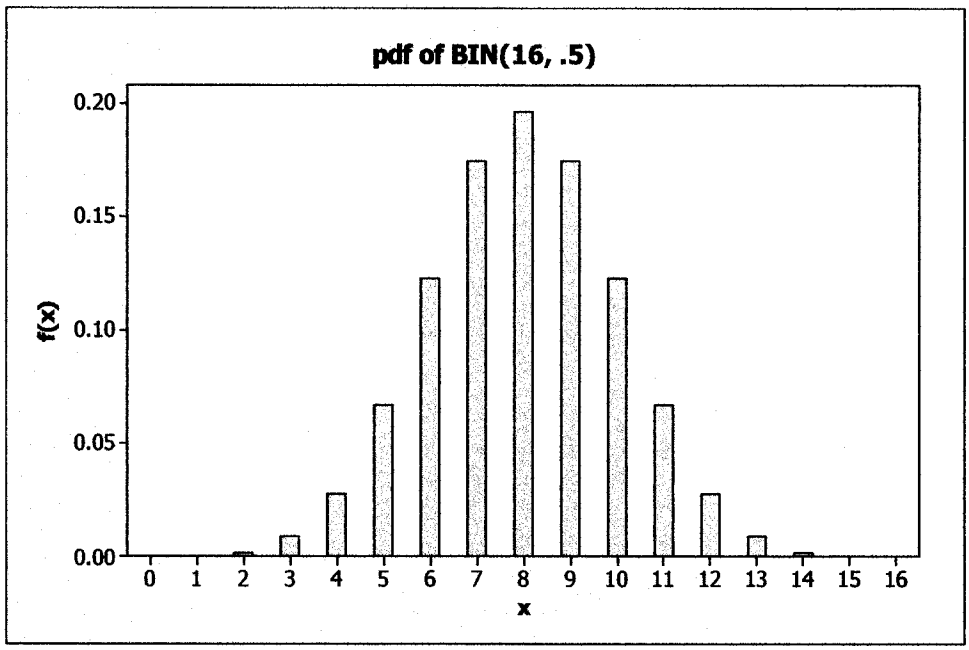

Figure 3: Graph of the binomial distribution BIN (n, p) for n = 16, p = .5

De Moivre (1733) figured that if he could approximate a mathematical expression for this curve, he would be able to solve problems such as finding the probability of 80 or more heads out of 200 coin flips much more easily. The curve he discovered is now called the normal distribution, and forms the basis of a large majority of statistical formulas. De Moivre's paper was discovered by Karl Pearson in 1924. Laplace (1783) used the normal curve to describe the distribution of errors. Gauss (1809) used it to analyze astronomical data. Due to the Central Limit Theorem, the normal distribution is the most important probability model in statistical computations.

Log-normal distribution:

Francis Galton presented the memoir of D. McAlister to the Royal Society of London (1879), according to which the log-normal distribution was introduced by D. McAlister, who derived the mean, the median, mode and the second moment of the distribution. In this presentation, Galton expressed the view that in certain situations, the geometric mean is a better measure of location than the arithmetic mean. Kapetyn (1903), the Dutch astronomer, described a mechanical device for generating samples from a log-normal population, similar to the mechanical device of Galton for generating normally distributed samples. The log-normal distribution has found applications in various branches of science:

Environmental Engineering: The probability distribution of contaminant concentrations is often modeled by the log-normal distribution (see, for example, Ott, 1978).

Ecology: The abundance of plant and animal species is quite often modeled by the log-normal distribution (see, for example, Sugihara, 1980; Magurran, 1988).

Geology and Mining: The probability distributions of concentrations of elements and their radioactivity have been modeled by the log-normal distribution (Ahrens, 1954).

Atmospheric Science: Many atmospheric physical and chemical properties are modeled by the log-normal distribution (Di Giorgio *et al.*, 1996).

A random variable X has a lognormal distribution if the random variable $Y = \ln X$ has a normal (i.e. Gaussian) distribution.

The normal distribution of Y is given by the density function:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

where μ is the mean, and σ is the standard deviation (σ^2 is the variance).

The density function of a lognormal distribution then becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{(\ln x-\mu)^2/2\sigma^2}$$

Note that the change in variables introduces an additional $\frac{1}{x}$ term outside of the exponential term. The corresponding complimentary cumulative distribution function for a lognormal distribution is given by:

$$\Pr[X \geq x] = \int_{z=x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma z} e^{(\ln z-\mu)^2/2\sigma^2} dz$$

The log-normal distribution with $\mu = 0$, $\sigma = 1$ is called Gibrat's distribution (Mansfield, 1962). It is known that the sum of two independent normal random variables Y_1 and Y_2 coming from an underlying normal distribution with means μ_1 and μ_2 and variances σ^2_1 and σ^2_2, is normal with mean $\mu_1+\mu_2$ and variance $\sigma^2_1+\sigma^2_2$. It follows that the product of two log normally distributed random variables also has a lognormal distribution.

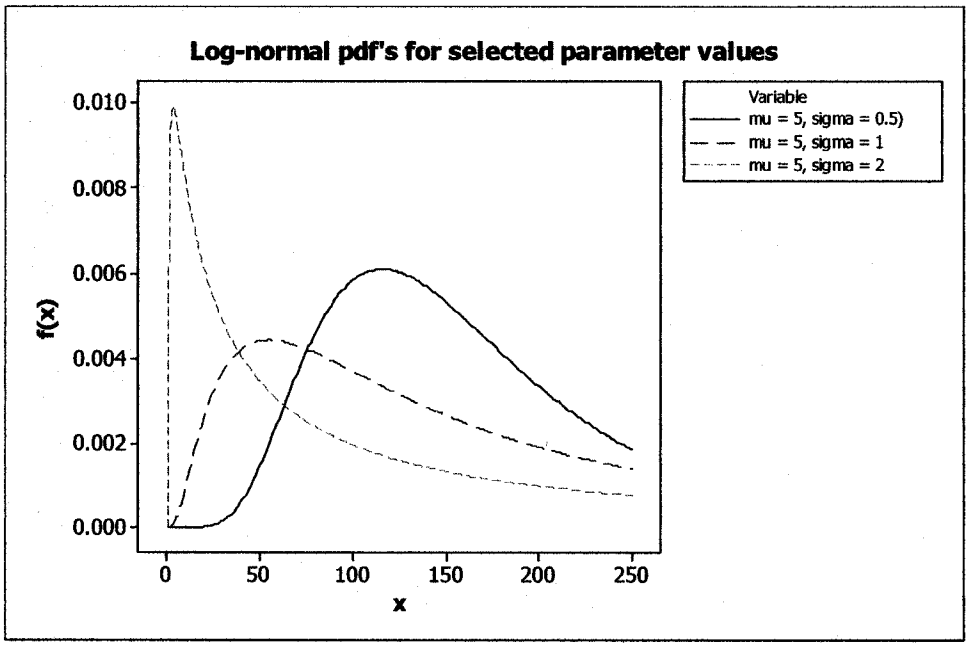

Figure 4: Graph of selected log-normal distributions

The parameters of interest of a lognormal distribution $LN(\mu,\sigma^2)$, are given below:

1. $Mean: \mu_1 = e^{\mu+0.5\sigma^2}$

2. $Median: e^{\mu}$

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

4. $CV: \frac{\sigma_1}{\mu_1} = \sqrt{(e^{\sigma^2} - 1)}$

5. $Skewness: \left(\frac{\sigma_1}{\mu_1}\right)^3 + 3\left(\frac{\sigma_1}{\mu_1}\right)$

1.1 Use of the Log-Normal Distribution in Environmental Statistics

It is clear from the above expressions for CV and skewness that the log-normal distribution is positively skewed, its skewness is a function of the parameter σ alone, and that the skewness increases with σ.

Contaminant concentration data from Superfund sites is quite often positively skewed (Singh *et al.,* 1997) and EPA guidance documents recommend using the log-normal distribution based formulas for computing the Upper Confidence Limits (UCL) for the mean contaminant concentration, or for the determination of number of samples for future sampling (Stewart, 1994). The log-normal distribution is very commonly used in environmental work, since it is very easy to use.

It has been pointed out in the statistical literature (Singh *et al.,* 1997), however, that (i) a normally distributed dataset with a few extreme observations on the high side can be incorrectly modeled by the log-normal distribution, and (ii) data from a site that has both low and high contaminant concentrations can also be incorrectly modeled by a log-normal distribution. This typically results in unreasonably high UCL values when the log-normal theory based H-statistics formula is used.

In this thesis, an attempt is made to demonstrate some of the problems one encounters by the use of such methods and the

unreasonable behavior of the log-normal theory based statistical procedures.

CHAPTER 2

COMPARISON OF SITE AND BACKGROUND DATA BASED ON THE LOG-NORMAL DISTRIBUTION

When a pollutant data set contains values that could be potential outliers, causing the data set to be skewed, taking the log transform masks those extreme points, which escape analysis when modeled and analyzed using lognormal distribution, as demonstrated by Example 2.1.

Example 2.1: Consider a simulated data set of 5 samples from a normal distribution with mean 50 and standard deviation 1.5 (background concentration) and a data set from a normal distribution with mean 150 and standard deviation 95 (contaminant concentration):

50.3499, 50.4863, 47.9185, 48.3566, 48.0776, 198.871, 224.345, 127.370, 13.8349, 114.570

This mixture of 10 samples has a mean of 92.4 and standard deviation 71.5. The data set is tested for normality (Figure 5-a) and then tested for log normality (Figure 5-b).

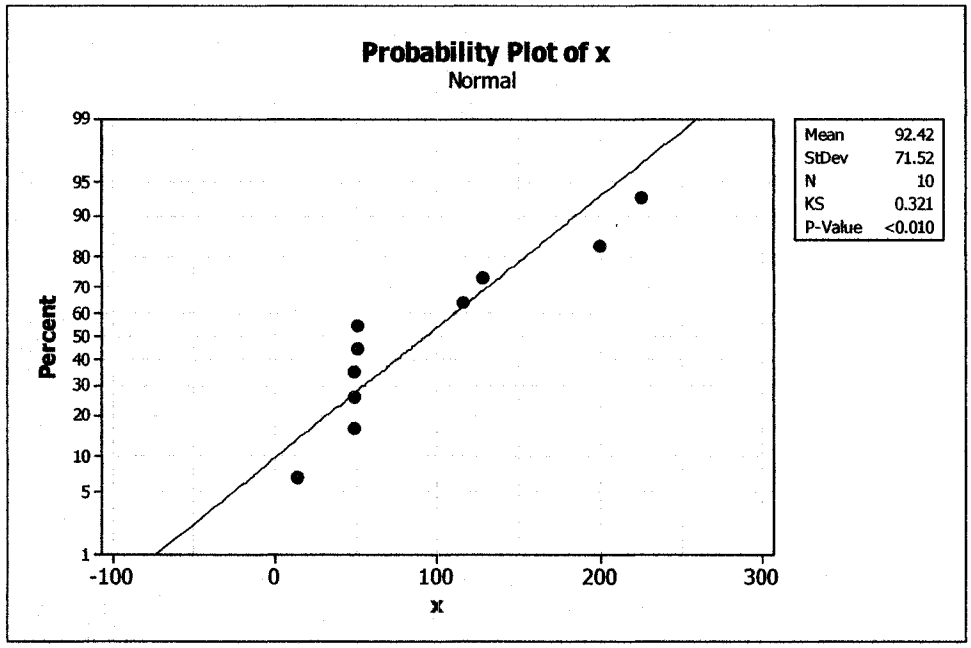

Figure 5-a: KS Test for Normality

The test rejects the null hypothesis of normality for this sample.

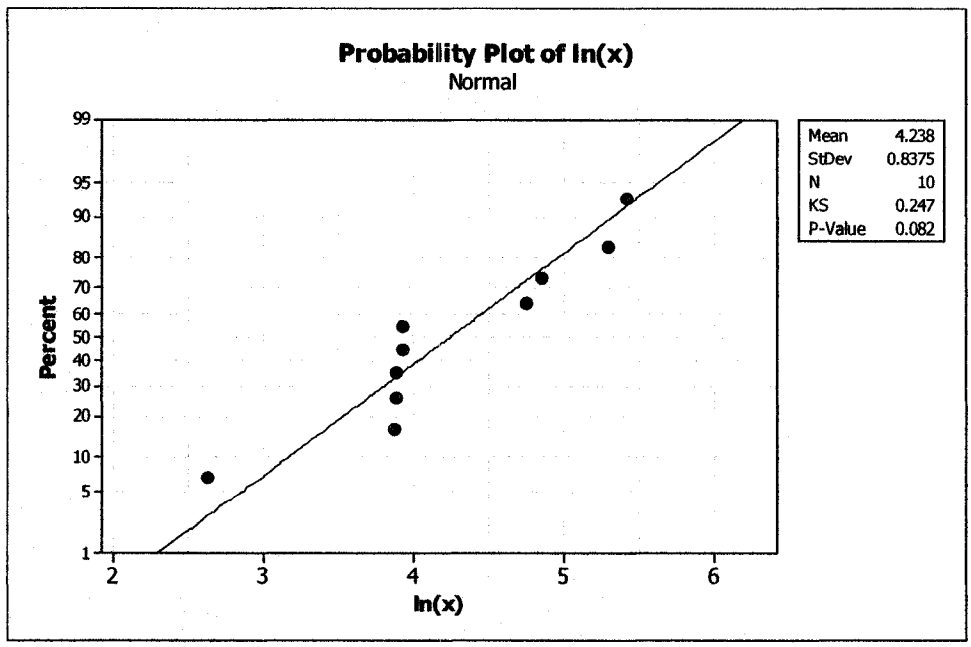

Figure 5-b: KS Test for Log Normality

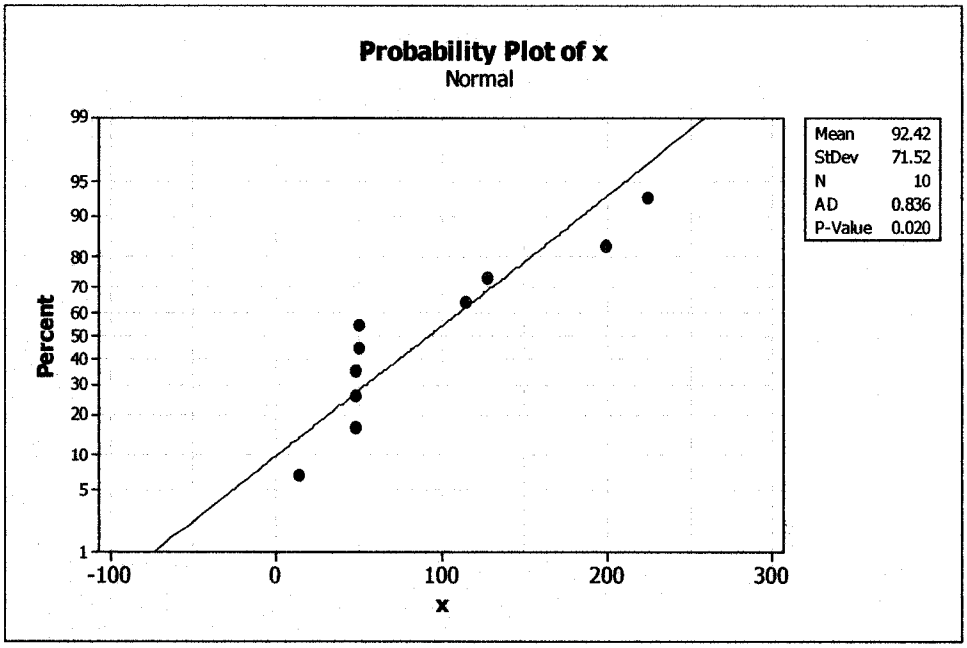

Figure 5-c: Anderson Darling Test for Normality

The test rejects the null hypothesis of normality for this sample.

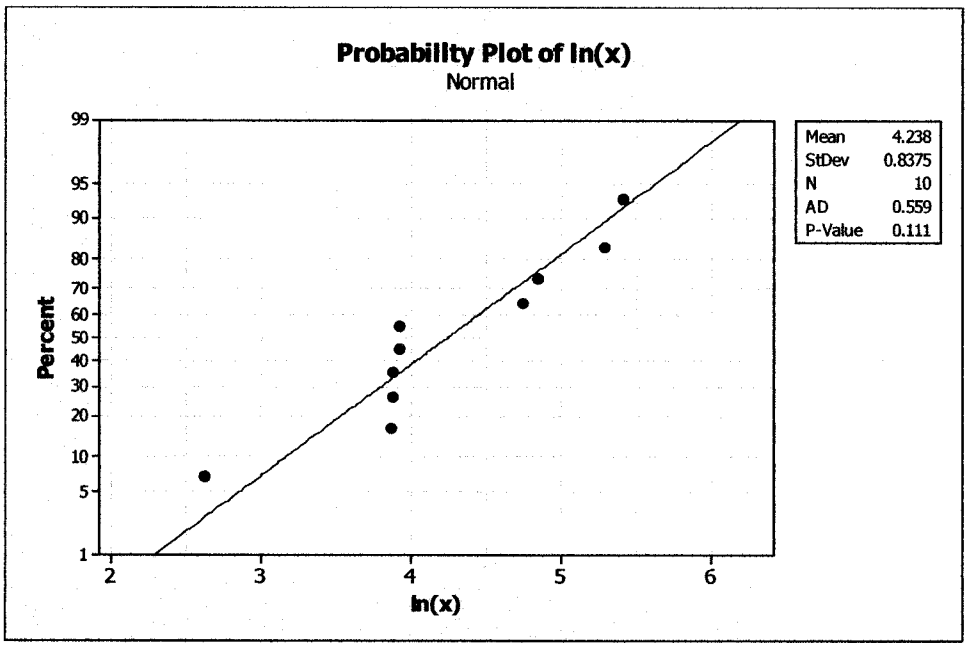

Figure 5-d: Anderson Darling Test for Log Normality

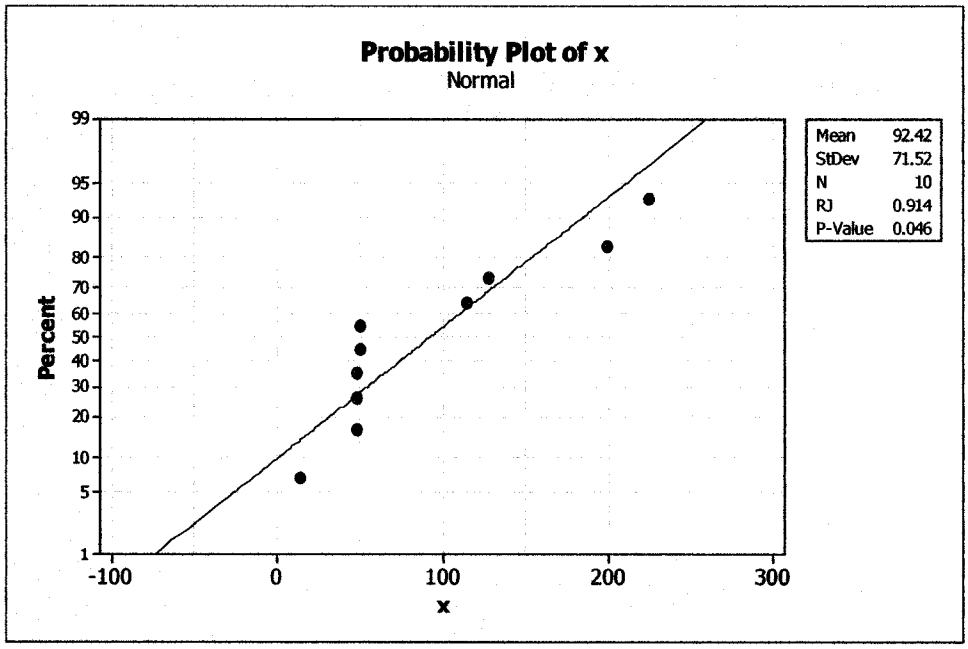

Figure 5-e: Ryan Joiner (similar to Shapiro Wilk) Test for Normality

The test rejects the null hypothesis of normality for this sample.

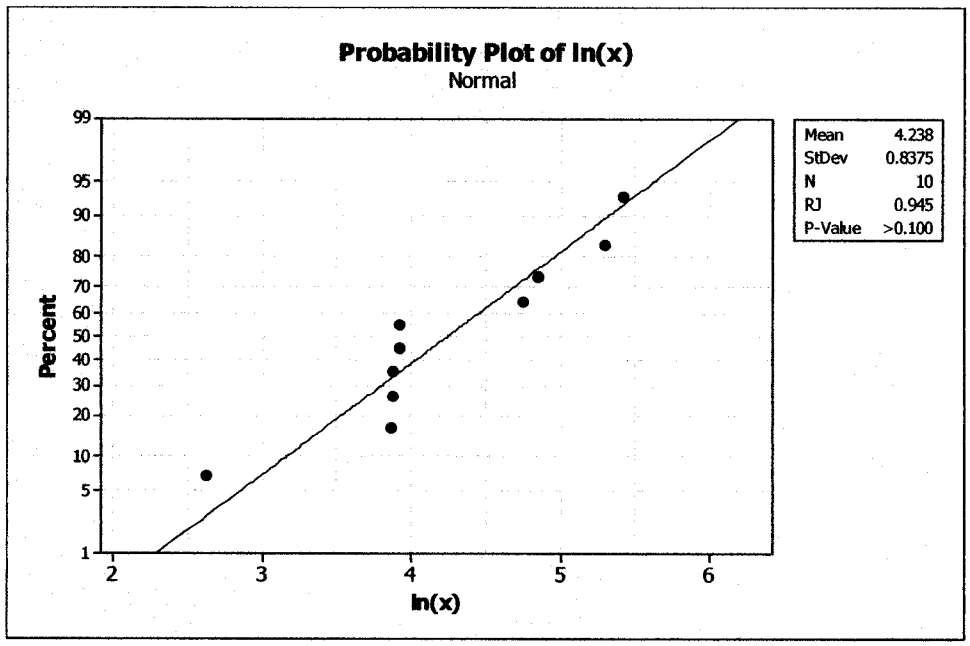

Figure 5-f: Ryan Joiner (similar to Shapiro Wilk) Test for Log Normality

In Example 2.2, two datasets simulating Background and Site conditions are generated.

Example 2.2: Background and Contaminated sites (simulated) data illustrating how taking the logarithm can lead to incorrect results:

- 20 data points each are generated from log-normal distributions (Background Data with mean = 5 and sd = 2 and Contaminated Data with mean = 5 and sd = 4). The true population means are 1096.6 (Background) and 442413.4 (Site).
- Their log transforms are taken and probability plots for each one of them are plotted (Figures 6-a and 6-d). The data clearly appears to be log-normally distributed.

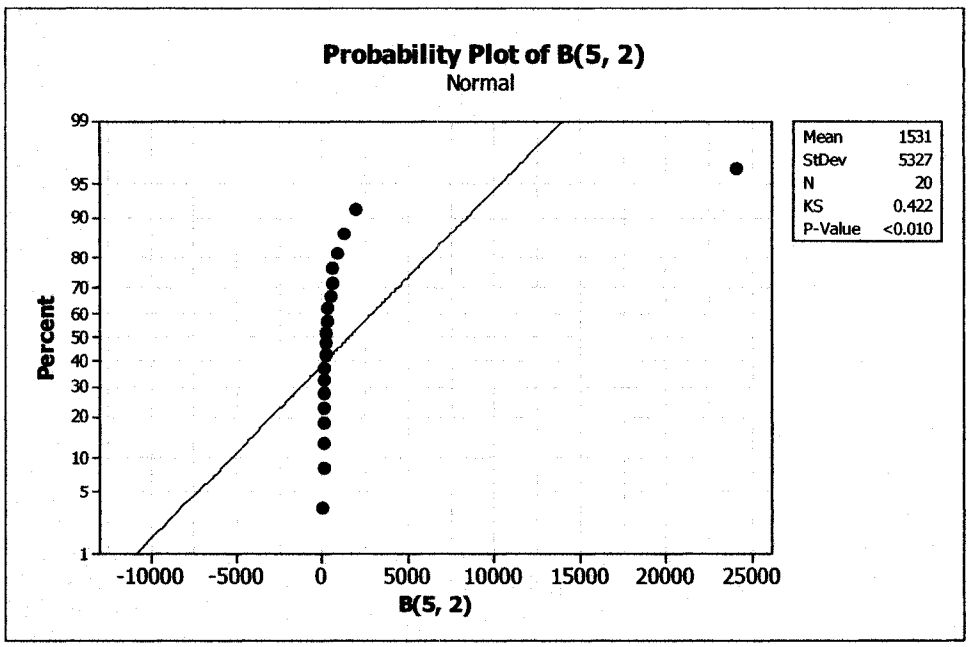

Figure 6-a: Test of Normality for Background Data

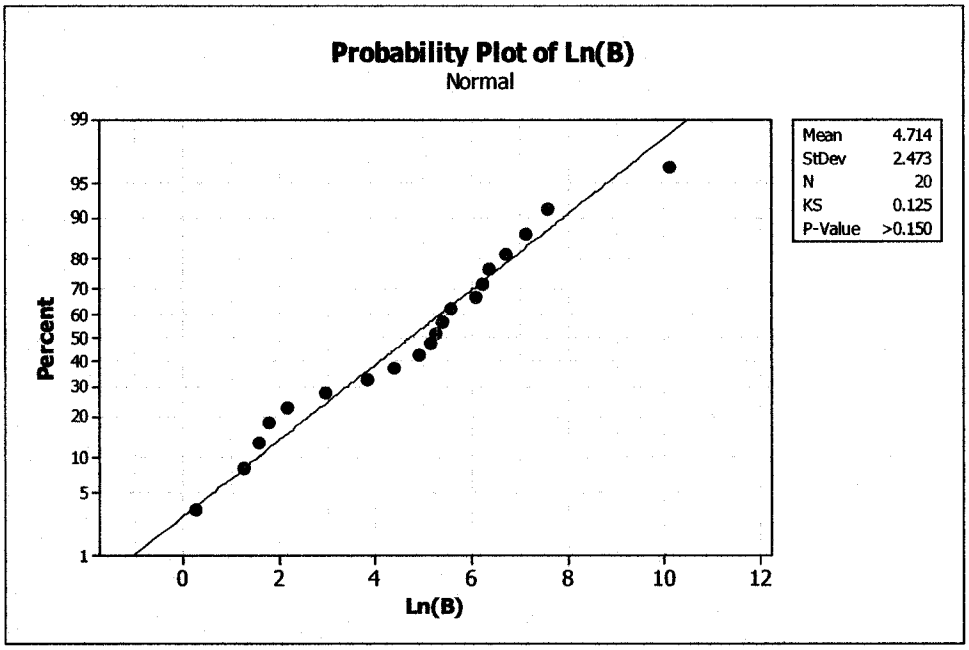

Figure 6-b: Test of Log-Normality for Background Data

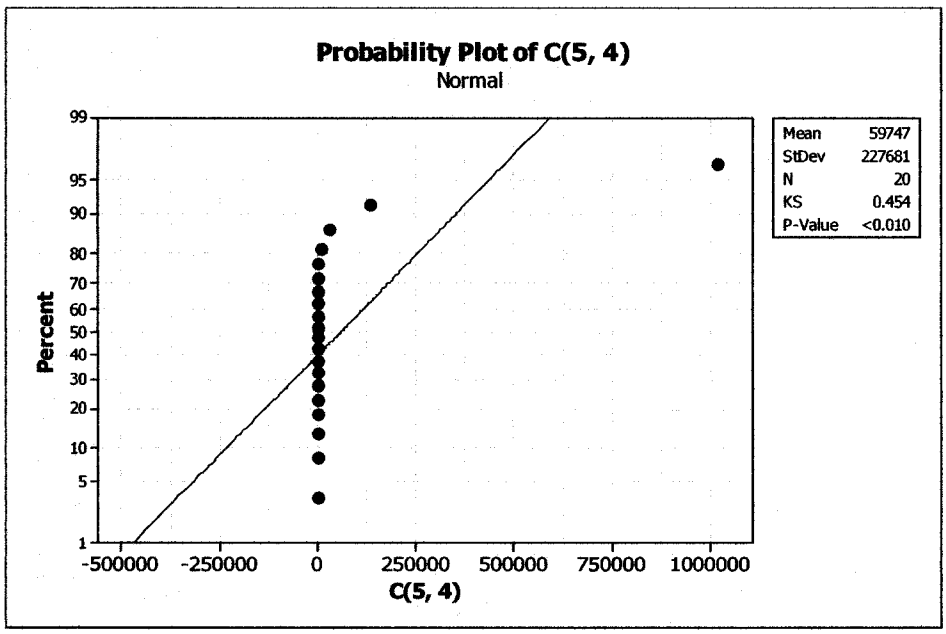

Figure 6-c: Test of Normality for Site Data

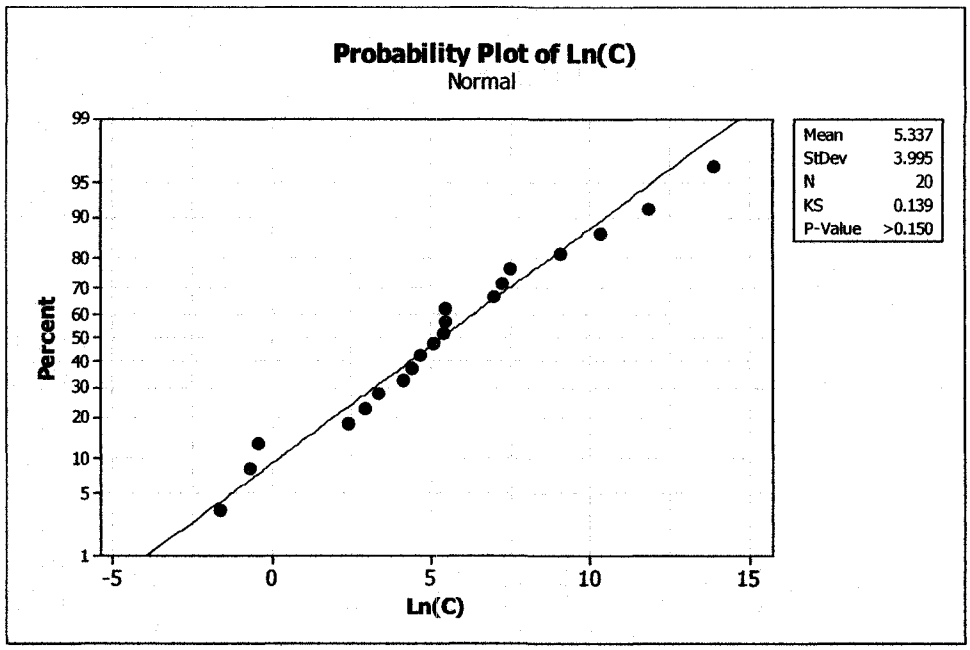

Figure 6-d: Test of Log-Normality for Site Data

Two-sample T-Test for Site vs. Background:

	N	Mean	StDev
Ln(B)	20	4.71	2.47
Ln(C)	20	5.34	3.99

T-Test of difference = 0 (vs not =): T-Value = -0.59 P-Value = 0.558 DF = 31

This dataset was generated with completely different background (B) and site (C) means, yet the 2-sample t-test on log-transformed data declared the two means to be equal in the log scale. In Chapter 3, we use Monte Carlo simulation to estimate the power of the 2-sample t-test based on the log-transformation of the Background and Site data.

CHAPTER 3

POWER OF THE 2-SAMPLE T-TEST BASED ON THE LOG-TRANSFORMATION

This section specifically deals with the following problem: Environmental engineers quite often use the 2-sample t – test on log-transformed data to compare Background and Site data, and many important decisions are made based on the conclusions from these tests. A study of the power of the t – test on raw as well as the transformed data sets has been carried out in this chapter.

In order to show that, when sample sizes are low to moderate (between 10 – 45) it is not possible to distinguish between log-normal and gamma distributions, the simulation in this chapter was done using the gamma distribution. In each instance, it was observed that the log-normal distribution fitted the sample generated from a gamma distribution. One example (Example 3.1) is included in the thesis.

Performing power analysis and sample size estimation is an important aspect of experimental design, because without these calculations, sample size may be too high or too low. If sample size is too low, the

experiment will lack the precision to provide reliable answers to the questions it is investigating. If sample size is too large, time and resources will be wasted, often for minimal gain. Therefore power calculations with different sample sizes and shape parameters were conducted, and for each set of parameters, a graph was plotted with power and difference in means as variables.

The methodology used in the thesis for estimating the power of the t-test is outlined below:

- Data sets from Site (Y) and Background (X) conditions were simulated from two gamma populations.
- Power of the T - test was estimated using Monte Carlo simulation for the raw samples and the log-transformed samples.

The programming for this part of the thesis was done in the programming language R.

Example 3.1: In this simulated example, one sample set of size 30 was drawn from G (shape = 2.5, scale = 1) representing Background (X), and another sample set of same size was drawn from G (shape = 2.0, scale = 1), representing Site (Y). Figures 7a-d show the results of testing normality and log-normality on the generated data sets. Both the Background and Site data turn out to be non-normal, and pass the test of log-normality.

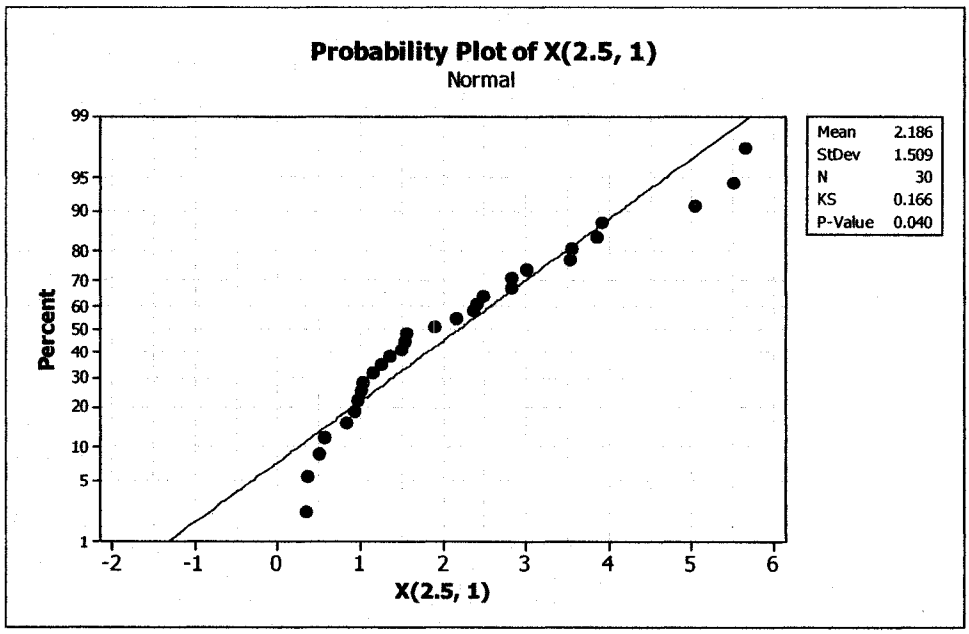

Figure 7-a: Test of Normality for Background

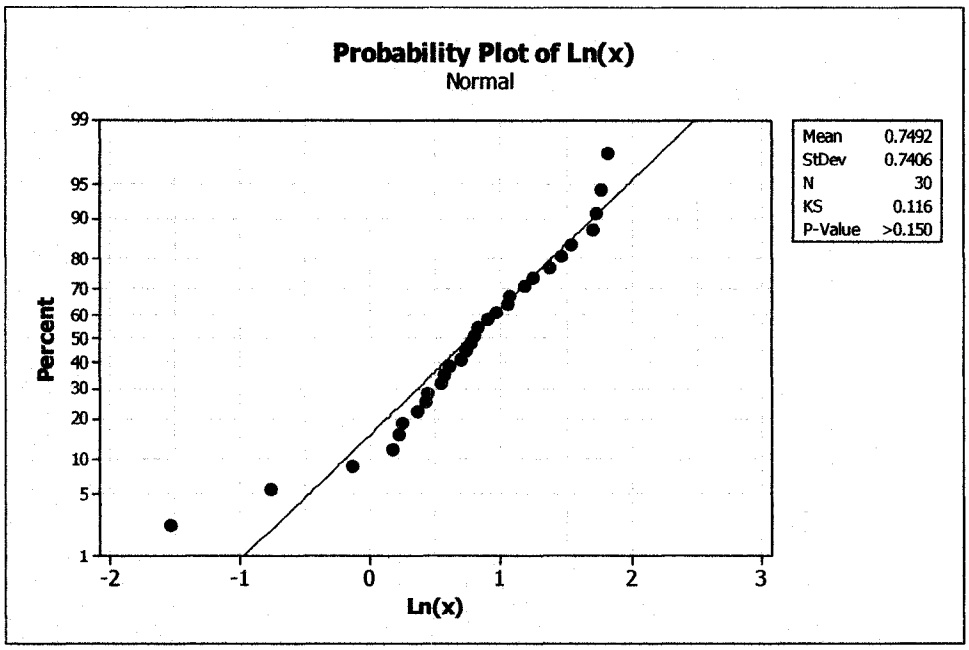

Figure 7-b: Test of Log-Normality for Background Data

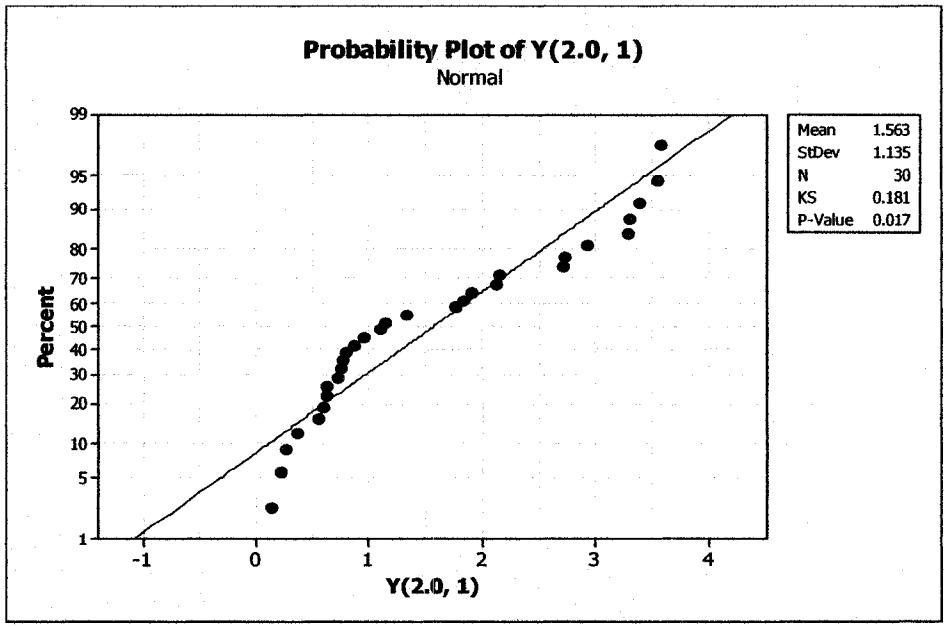

Figure 7-c: Test of Normality for Site Data

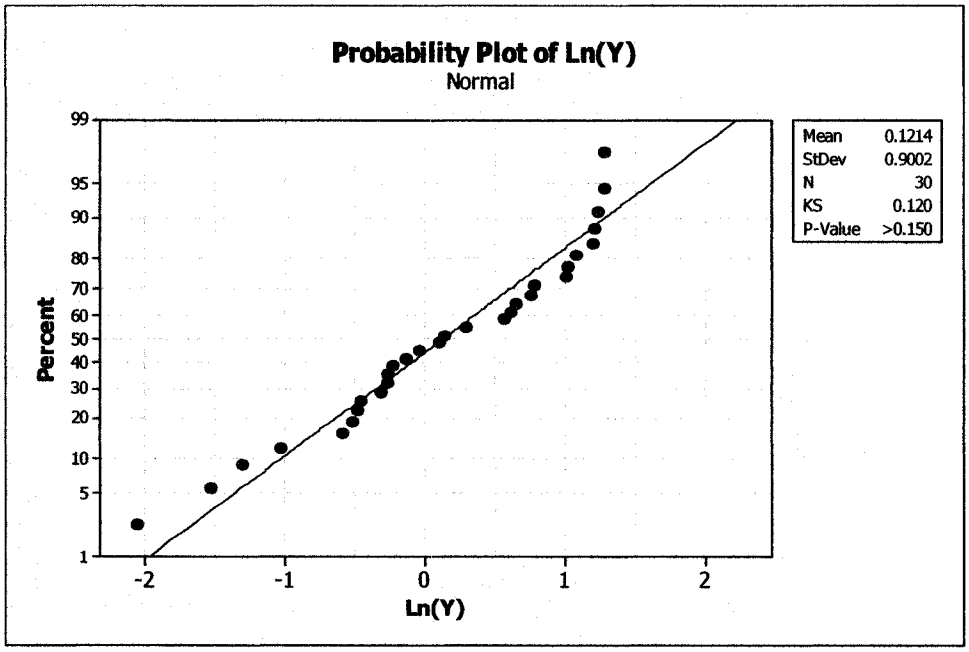

Figure 7-d: Test of Log-Normality for Site Data

3.1 Comparison of Powers of T-Tests Based On Raw and Log - Transformed Data

In order to estimate the power of the 2-sample t-test, 2-sample data were generated from the gamma distributions GAM (α_1, β_1) and GAM (α_2, β_2) with varying values of the difference in means $\alpha_1\beta_1 - \alpha_2\beta_2$.

The values of the shape parameter were chosen so that skewness for the first sample was 1.265, and the skewness for the second sample ranged from 0.7727 to 1.265:

$$Skewness = \frac{2}{\sqrt{\alpha}} = \frac{2}{\sqrt{2.5}} = 1.265$$ (Skewness kept at 1.265 throughout under X)

$$Skewness = \frac{2}{\sqrt{\alpha}} = \frac{2}{\sqrt{6.7}} = 0.7727$$ (Skewness ranges from 1.265 to 0.7727 under Y)

Steps of the simulation experiment to estimate the power are given below:

1) Generate $x_1, x_2, \ldots, x_n \sim GAM(\alpha_1, \beta_1)$, $y_1, y_2, \ldots, y_n \sim GAM(\alpha_2, \beta_2)$.
2) Run the 2-sample t-test for unequal variances on the two samples.
3) Repeat Steps 1-2 N times (N large integer), and count the number of times the null hypothesis of equal means is rejected.
4) Estimate power as follows:

Power = # of rejections/N

The generated data and the complete outputs from ProUCL are included in Appendix C. Tables 1 – 6 (Appendix A) show the power function of the 2 – sample t – tests performed on raw and log – transformed data, computed in R. Figures 8 – 13 show the estimated power function of the two t – test procedures.

GRAPHS

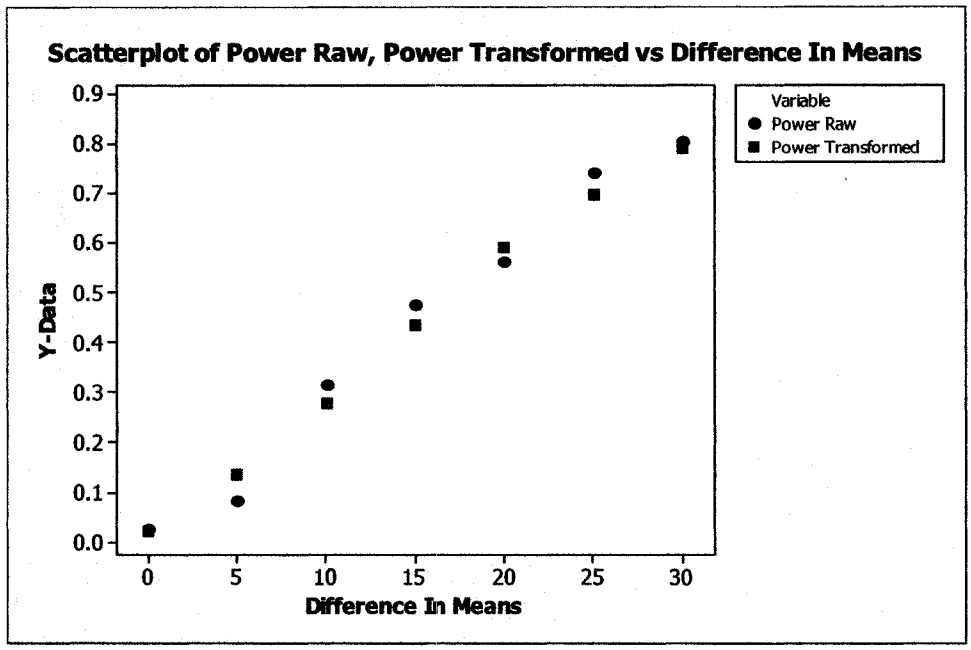

Figure 8: n = 5, X ~ G (0.5, 10) vs. Y ~ G (0.5...3.5, 10)

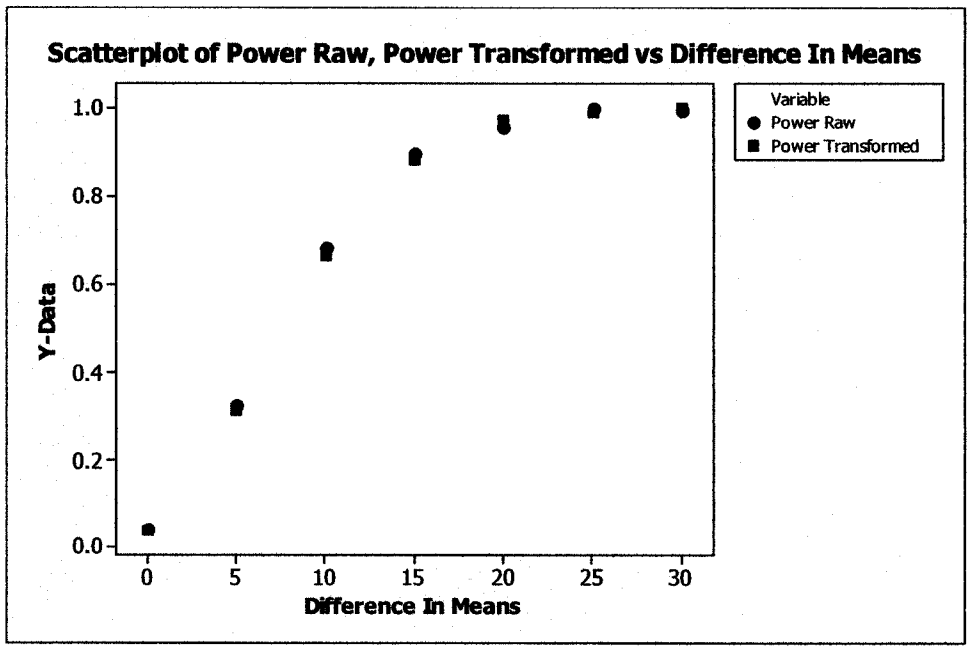

Figure 9: n = 10, X ~ G (0.5, 10) vs. Y ~ G (0.5...3.5, 10)

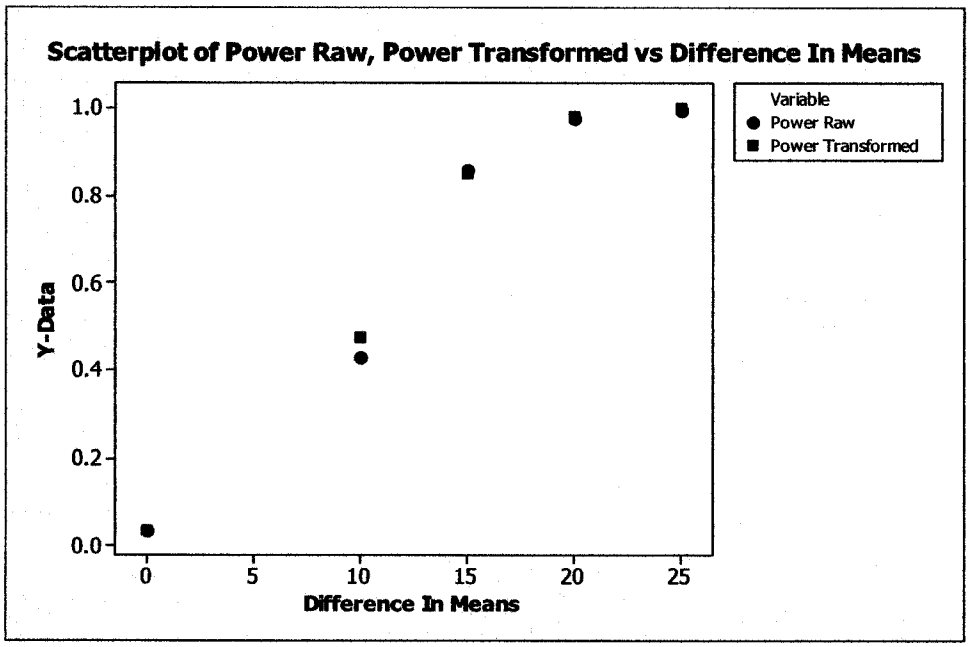

Figure 10: n = 15, X ~ G (0.5, 10) vs. Y ~ G (0.5...3.5, 10)

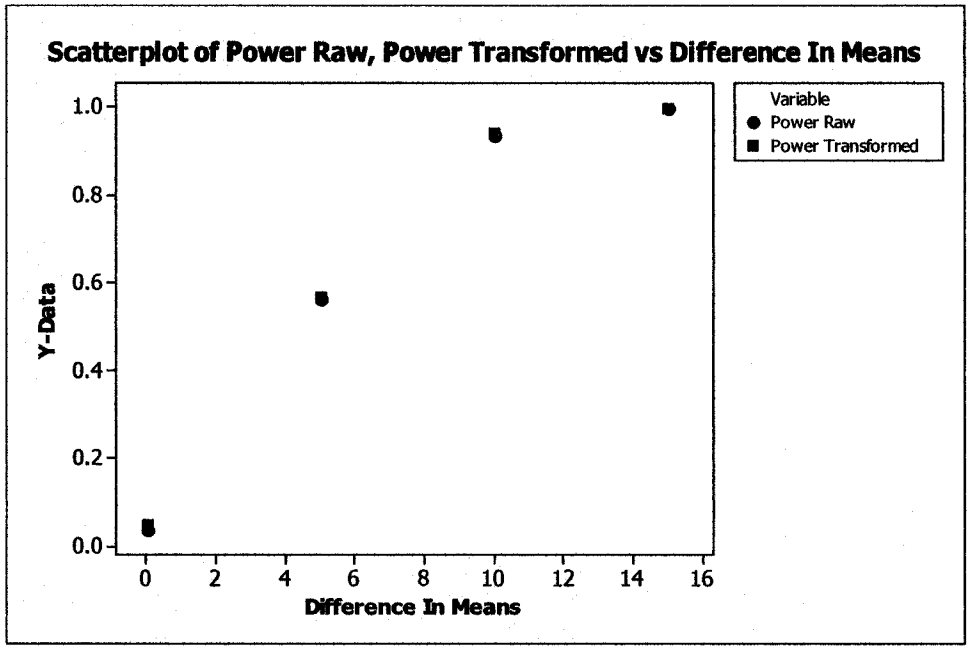

Figure 11: n = 20, X ~ G (0.5, 10) vs. Y ~ G (0.5...2.0, 10)

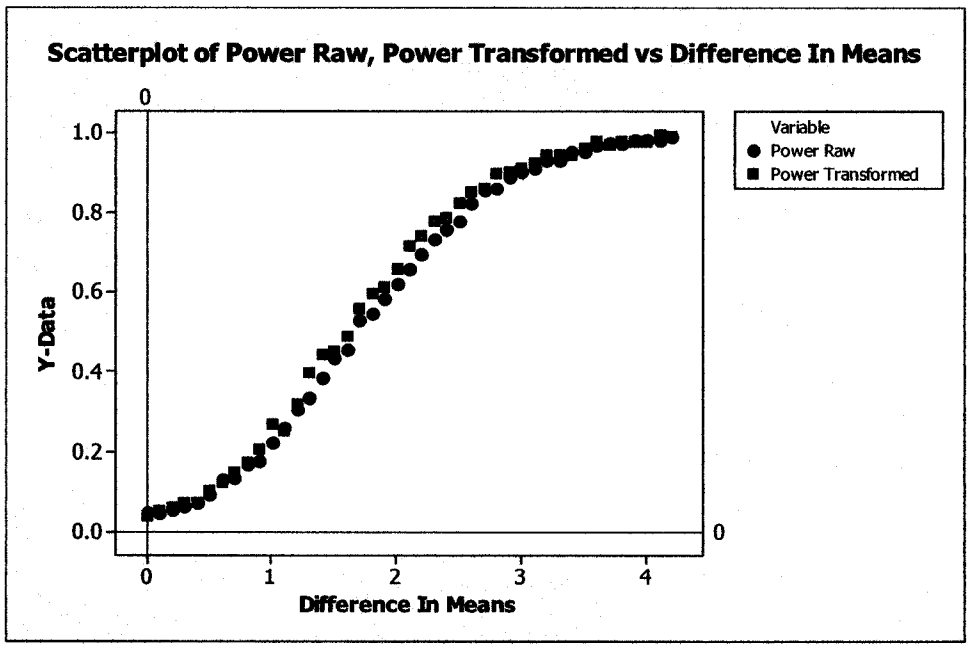

Figure 12: n = 10, X ~ G (2.5, 1) vs. Y ~ G (2.5...6.7, 1)

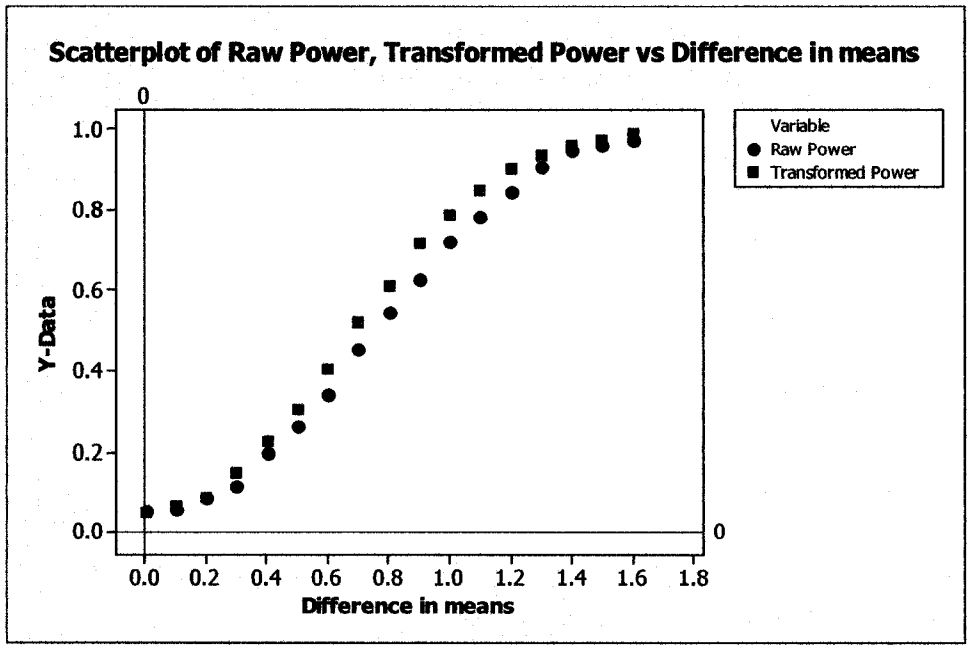

Figure 13: n = 40, X ~ G (2.5, 1) vs. Y ~ G (2.5...3.6, 1)

3.2 Discussion of Results

From the above study, it is clearly seen that as the difference in means increases, the power of the t-tests based on both the raw and the transformed data increases, as expected. It is also observed that the power of the t-test based on the raw data is nearly the same as the power of the t-test based on the log - transformed data.

This shows that taking the log transform is not really necessary. Use of lognormal distribution in modeling environmental data has come under extensive criticism by many authors; Singh and Nocerino (1995) have shown that when dealing with positively skewed data, non parametric methods give more reliable estimates of the population.

As studied by Staudte and Sheather (1990), the tests based on the Student's t are non robust in the presence of outliers. Singh, Singh, and Engelhardt (1997) also have shown that the log normal distribution could be deceptive as it often hides the outliers.

CHAPTER 4

PERFORMANCE OF H – UCL IN PRESENCE OF NON DETECTS

Censored data occurs in environmental studies when pollutant levels fall below the detection (or reporting) limits of instrumentation. Estimation of population parameters or testing hypotheses from censored data sets are problematic (see Helsel, 2005, or Hinton, 1993).

The problem of non-detects (also called left censoring) occurs commonly in environmental data. A "non-detect" is an observation that is below the limit of detection of an analytical method. The limit of detection is generally defined as the lowest concentration that can be determined to be statistically different from a blank specimen. The limit of detection is an imprecise quantity that can vary from sample to sample and laboratory to laboratory. The most common method of dealing with non-detects in environmental samples is the substitution method, in which the values below detection limit (DL) are replaced by 0, DL/2, or DL.

As mentioned earlier, contaminant concentration data sets from Superfund sites are typically positively skewed, and EPA Guidance Documents (such as USEPA, 1987) recommend the use of H-statistic

based Upper Confidence Limits (UCL) for the mean, which is based upon log-normal theory:

$$H-UCL = e^{\bar{y}+0.5s_y^2+s_y H_{1-\alpha}/\sqrt{n-1}}$$

where

$y_i = \ln(x_i)$ = log-transformed concentration

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

and $H_{1-\alpha}$ values are the upper % - points of Land's H - Statistics (Land, 1975 or Gilbert, 1987).

The behavior of the H-statistic based UCL when there are non-detects in the sample has not been investigated in environmental statistics literature. In this chapter, we simulate samples with varying proportions of non-detects, and compute the H-statistic based 95% UCL for the mean using the three substitutions. The simulation experiment used in the thesis is outlined below:

1. Generate a sample of size n (n = 10, 50, and 100) from $LN(\mu,\sigma)$ in MINITAB. The parameters of the log-normal distribution were chosen so as to simulate datasets with (a) low skewness, and (b) high skewness.

2. A detection limit (DL) was chosen for a generated (complete) sample so that p% of the observations are '< DL', for p = 10, 20, 30, and 40.
3. The software package ProUCL was then used to compute the H-UCL of the mean for the full data, and also the datasets obtained from the three substitution methods.

Low skewness: $\mu = 2$, $\sigma = 0.5$

Mean = 8.37

$$CV = \sqrt{\exp(\sigma^2 - 1)}$$
$$= 0.5329$$

$$Skewness = (CV)^3 + 3(CV)$$
$$= 1.75$$

Data sets of sizes n = 10, 50, and 100 were generated. These data sets are included in Appendix C of this thesis, along with complete outputs obtained from ProUCL. The results are summarized in Tables 7 – 10 below.

High skewness: $\mu = 2$, $\sigma = 2.5$

Mean = 168.17

$$CV = \sqrt{\exp(\sigma^2 - 1)}$$
$$= 13.805$$

$$Skewness = (CV)^3 + 3(CV)$$
$$= 2672.105$$

Data sets of sizes n = 10, 50, and 100 were generated. These data sets are included in Appendix B of this thesis, along with complete outputs obtained from ProUCL. The results are summarized in Tables 10 – 12 below.

4.1 Discussion of Results

It can be seen from Tables 7 – 9 (Appendix B) that (i) when skewness is low and n is small (10), substitution of '<DL' values by 0 inflates the H-UCL quite a bit, but the other two substitution methods work reasonably well. When skewness is high (Tables 10 – 12, Appendix B), and sample size is low (n = 10), the H-UCL obtained from any of the substitution methods is orders of magnitude higher than the true mean. The situation improves a bit for moderate (n=50) and large (n=100) sample sizes, but the H – UCL of the censored data is still unreasonably high.

It should be kept in mind that when an observation in a sample is replaced by a smaller value, the sample mean is going to decrease, yet the H-UCL goes sky-high in some of the examples presented here.

REFERENCES

1. U.S. Environmental Protection Agency (1987). Methods for Evaluating the Attainment of Cleanup Standards. Vol. 1: Soils and Soil Media (EPA 230/02-89/042).

2. L. H. Ahrens (1954), The log-normal distribution of the elements. (A fundamental law of geochemistry and its subsidiary).

3. C. Di Giorgio, A. Krempkoff, H. Giraud, P. Binder, C. Tiret, G. Dumenil (1996), Atmospheric pollution by airborne microorganisms in the City of Marseilles. Atmospheric Environment, vol. 30, pp. 155-160.

4. F. Galton (1879). The Geometric Mean in Vital and Social Statistics. Proceedings of the Royal Society, v. 29, p. 367.

5. Steven W. Hinton (1993), Delta Lognormal Statistical Methodology Performance. Env Science Tech, Vol. 27, No. 10.

6. J. C. Kapetyn (1903), Skew Frequency Curves in Biology and Statistics. Astronomical Laboratory, Groningen, Noordhoff. Geochimica et Cosmochimics Acta 5: pp. 49-73.

7. C.E. Land (195), Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance, in Selected Tables in Mathematical Statistics, vol. III, American Mathematical Scoiety, Providence, R.I., 385 – 419.

8. A. E. Magurran (1988), Ecological Diversity and its Measurement. London. Croom Helm.

9. E. Mansfield (1962), Entry, Gibrat's Law, Innovation, and The Growth of Firms. American Economic Review, pp. 1023-1051.

10. W. R. Ott (1978). Environmental Indices. Ann Arbor, MI, Ann Arbor Science.

11. Ashok K. Singh, Anita Singh, and M Engelhardt, Some Practical Aspects of Sample Size and Power Computations for Estimating the

Mean of Positively Skewed Distributions in Environmental Applications, EPA/600/s-99/006, November 1999.

12. Ashok K. Singh, Anita Singh, and M. Engelhardt, The Log-normal Distribution in Environmental Distributions. EPA/600/R-97/006, December 1997.

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

14. Sally L. Stewart (1994), Use of Log normal transformations in Environmental Statistics, M.S. Thesis, Department of Mathematics, University of Nevada, Las Vegas.

15. G. Sugiharra (1980), Minimal community structure: An explanation of species abundance patterns, American Naturalist Vol. 116: pp. 770-786.

VITA

Graduate College
University of Nevada, Las Vegas

Devarshi Pant

Home Address:
1555, East Rochelle Avenue,
171, Las Vegas, NV 89119

Degrees:
Bachelors in Environmental Engineering, 2001
Shivaji University, India

Thesis Title: On the Use of Lognormal Distribution in Environmental Data Analysis

Thesis Examination Committee:
Chairperson, Dr. Ashok Kumar Singh, Ph.D.
Committee Member, Dr. Deudonne D. Phanord, Ph.D.
Committee Member, Dr. Rohan Dalpatadu, Ph.D.
Graduate Faculty Representative, Dr. Laxmi P. Gewali, Ph.D.