1-1-2005

An investigation of the Kaplan-Meier Upper Confidence Limit for the population mean from environmental samples with nondetects

Violeta Graciela Hennessey
University of Nevada, Las Vegas

AN INVESTIGATION OF THE KAPLAN-MEIER UPPER CONFIDENCE LIMIT FOR THE POPULATION MEAN FROM ENVIRONMENTAL SAMPLES WITH NONDETECTS

by

Violeta Graciela Hennessey

Bachelor of Science
Texas State University
2003

A thesis submitted in partial fulfillment
of the requirements for the

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

Graduate College
University of Nevada, Las Vegas
August 2005

UMI Number: 1429707

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

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval
The Graduate College
University of Nevada, Las Vegas

July 22, 2005

The Thesis prepared by

Violeta G. Hennessey

Entitled

An Investigation of the Kaplan-Meier Upper Confidence Limit for the Population Mean From Environmental Samples with Nondetects

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

Master of Science in Mathematical Sciences

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

1017-53

ABSTRACT

An Investigation of the Kaplan-Meier Upper Confidence Limit for the Population Mean from Environmental Samples with Nondetects

by

Violeta Graciela Hennessey

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

The Kaplan-Meier (K-M) estimator is a non-parametric estimator of the survival function, used in lifetesting and medical follow-up studies where some of the observations are incomplete (right-censored data). In environmental applications, the user is faced with the problem of contaminant concentration falling below the limit of detection (DL) of the instrument (left-censored data). The K-M estimator has recently been proposed in environmental literature for computing the Upper Confidence Limit (UCL) of the mean in the presence of nondetects in environmental data sets. The properties of this UCL, however, have not been investigated. In this thesis, I propose to use Monte Carlo simulation to study the performance of the K-M method for computing the UCL of the mean.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

A special thanks to my advisor, Dr. Ashok K. Singh, for all of his patience and help in completing this thesis. Thanks to my professors, Dr. Dalpatadu, Dr. Ananda, Dr. Murphy, and Dr. Ho for their knowledge and support during my educational career at UNLV. I would also like to say thank you to my husband, Youssef, for continuously pushing me to work on my thesis and my mother, Maria, for putting up with my endless college life. I dedicate this thesis to my late father, Michael Hennessey.

CHAPTER 1

INTRODUCTION

Standard statistical analysis of data starts with the assumption of normality, but environmental data sets are typically positively skewed [2]. The situation gets even more complicated when the contaminant concentration data has nondetects. This happens when the concentration of a contaminant is below the detection limit (DL) of the analytical instrument [7]. The problem of nondetects in an environmental sample occurs quite frequently. An environmental data set that contains nondetects is referred to as censored data. When the environmental data set contains measurements that fall below the DL, the data set is referred to as left-censored [7]. So how does a scientist deal with these measurements that are below the DL?

It is traditional for environmental scientists to use the substitution method. The substitution method replaces the value observed below the DL with a value of zero, or a value of one-half the detection limit (DL/2), or by DL itself in order to create uncensored data for ease of statistical analysis [7]. Replacement by 0 results in a biased low mean and a biased high standard deviation. Replacement by DL results in a biased high mean and a biased low standard deviation. In some applications, the United States Environmental Protection Agency (EPA) does not even require that the values below the DL be reported [7], but this may lead to biased estimates and may not be protective of the environment. These methods can also create unnecessary expenditures for the Potentially

Responsible Party (PRP) when, for example, a site is declared unclean when it is actually clean.

The problem of nondetects occurs in life-testing and medical follow-up studies as well, but the nondetects occur in a different manner. It is common for the data to be right-censored, containing observations that fall above a given value. What is being observed is usually a measurement of time, a nonnegative value [1]. Methods for right-censored data that do not ignore or substitute false values into the data set have been developed and deployed successfully in this field. One of the methods developed for right-censored data is the Kaplan Meier (K-M) estimator of the survival function, which can also be used to estimate the population mean [8]. This will be discussed in depth in Chapter 2 and its performance on left-censored data is the main objective of this thesis.

It has been proposed by Dennis Helsel to use the K-M method on censored environmental data (left-censored) for estimation of summary statistics for any size data set as long as the percentage of nondetects is less than 50% [3].

The purpose of this thesis is to investigate the performance of the K-M method for computing the Upper Confidence Limit (UCL) for the population mean from environmental samples with nondetects. This will be accomplished by using a Monte Carlo simulation experiment that incorporates the bootstrap method. The simulation experiment was implemented using SAS software on a Windows platform and a SAS source code designed specifically for this thesis experiment. The simulation experimented is discussed in detail in Chapter 3. The results are presented in Chapter 4, in which Minitab was used to generate the graphs. After analyzing and interpreting the results, conclusions were made and are presented in Chapter 5.

1.1 Terminology

The following terminology will be used throughout this thesis:

1. Bootstrap Method: is a method that incorporates sampling with replacement from a given sample for estimating specific statistics.
2. Censored Data: a data set that contains observations whose measurements are less than or greater than a given constant.
3. Detection Limit: the lowest value that a measurement can be in order to be detected with a reasonable degree of accuracy.
4. Left-Censored Data: a data set that contains a percentage of observations whose measurements are less than DL.
5. Nondetects: measurements that do not meet the criteria of being detected; observations that fall below DL.
6. Nonparametric Method: methods that do not require an assumption about the parametric form of the distribution of the data.
7. Right-Censored Data: a data set that contains a percentage of observations whose measurements are greater than a given constant.
8. Skewness: a measure of the asymmetry of a probability distribution.
9. Upper Confidence Limit: a value U, such that $P(\mu < U) = 1 - \alpha$, where $100(1 - \alpha)\%$ is the confidence level.

CHAPTER 2

THE KAPLAN-MEIER METHOD

2.1 Lifetesting

In lifetesting and medical follow-up studies, data is often incomplete (censored). What is being observed is the time that an event occurs, the event being death or failure [4]. For many reasons, studies are for a fixed period of time, where the start and end time is often set in advance. Right censoring occurs when a patient is lost to follow-up, when a patient is still alive at the end of a study, and when a patient dies of other causes [1]. All that is known is that the event of interest is greater than the observed time. In these cases the subject is considered a censored observation.

To understand this better, a scenario is presented. A researcher in a hospital observes the time of death in days of 10 patients who have been diagnosed with a terminal disease. One of the patients leaves the hospital at day 7 and contact is lost. The event of interest is death and since the death of the patient is only known to be greater than 7 days, day 7 is recorded and is labeled right-censored. At the end of the 30 day study, the day of death of 7 patients were recorded and 2 patients are still alive. How does the researcher record the day of death for the 2 patients that are still alive at the end of the study? What is common is to record day 30 for both and label them as right-censored [1].

Table 2.1 Example of a Recorded Lifetesting Data Set

Patient id	Time of death (t)	Censor (1 = uncensored, 0 = censored)
01	11	1
02	4	1
03	22	1
04	30	0
05	13	1
06	21	1
07	30	0
08	25	1
09	7	0
010	13	1

2.2 Survival Analysis

The statistical analysis of lifetime data is called survival analysis [8]. The lifetime data involves a nonnegative random variable T, often representing the lifetimes of that which is being observed from a known start time [1]. Any random variable that contains observations that lie in the interval $[0, \infty)$ can be considered a survival random variable [8]. In this thesis we will assume that T is continuous. Given that T has a probability density function (p.d.f.) f(t), the cumulative distribution function (c.d.f.) F(t), is defined as

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)\,du \,.$$

The survival function S(t) gives the probability of an individual surviving beyond time t, where S(0) = 1 and $S(\infty) = 0$. S(t) is a monotone nonincreasing continuous function given by,

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_t^\infty f(u)du$$

The mean or expected value of T is computed by calculating the area underneath the survival curve [1].

$$\mu_T = \int_0^\infty S(t)dt$$

2.3 Kaplan-Meier (K-M) Estimator

Censored data cannot be analyzed by using standard statistical methods. Survival analysis contains techniques that were developed for right-censored data. The most popular technique is the K-M estimator, also known as the product-limit estimator, was developed in 1958 by E.L. Kaplan and Paul Meier [4]. Given the survival random variable T, whose observations are nonnegative and that may contain right-censored observations, K-M is a nonparametric estimator of the survival function S(t) [1].

In environmental data sets, the contaminant concentration is a nonnegative random variable (X) that may contain left-censored observations. Table 2.2 is an actual environmental data set that contains nondetects (left-censored observations). This data set actually contains two DLs of $DL = 0.31$ and $DL = 0.10$. A censor variable is created that has a value of 1 if the observation is detectable (uncensored) and 0 if the observation falls below the DL (censored).

Table 2.3(a) Actual Left-Censored Environmental Data Set

i	X (x_i)	Censor (1 = uncensored, 0 = censored)
1	1.30	1
2	1.10	1
3	0.80	1
4	0.70	1
5	0.70	1
6	0.40	1
7	< 0.31	0
8	0.26	1
9	0.20	1
10	< 0.10	0
11	< 0.10	0

In order to use the K-M method to estimate S(x), the probability of a measurement being greater than x, X must be transformed to a right-censored data set (Y). This is accomplished by taking the maximum observation value (M = 1.30), and adding a reasonable chosen value ($\varepsilon = 0.30$) to it. Given that $L = M + \varepsilon$, take each observation (x_i) and subtract it from L ($y_i = L - x_i$) [3]. The following table shows the results of this observation.

Table 2.3(b) Left-Censored Data Set Transformed to Right Censored

i	Y ($y_i = 1.6 - x_i$)	Censor (1 = uncensored, 0 = censored)
1	0.30	1
2	0.50	1
3	0.80	1
4	0.90	1
5	0.90	1
6	1.20	1
7	1.29	0
8	1.34	1
9	1.40	1
10	1.50	0
11	1.50	0

2.3.1 K-M Implementation

The data set presented in Table 2.3(b) will be used to show the implementation of the K-M method [4]. The sample size of the data set is N = 11. The N observations (y_i) will be put into ascending order so that $0 \leq y_1' \leq y_2' \ldots\ldots \leq y_N'$. The measurement scale is divided into chosen intervals, $(0, u_1)$, (u_1, u_2), ... The chosen intervals are given in Table 2.3.1. Let n_i, δ_i, λ_i, and $\hat{S}(y)$ be defined as,

n_i = the number of observation at the beginning of the interval.

δ_i = the number uncensored observations that have occurred within the interval.

λ_i = the number of censored observations that have occurred within the interval.

$\hat{S}(y)$ = the estimated probability of a measurement being greater than $y = u_i$.

Given the information above, the implementation of the K-M method to estimate S(y) is shown in Table 2.4. The estimated survival curve is constructed with the computation of $\hat{S}(y)$. Figure 2.3.2 shows the estimated survival curve of variable Y.

Table 2.3.1 K-M Implementation Table

i	u_i	n_i ($n_{i+1} = n_i - \delta_i - \lambda_i$)	δ_i	λ_i	n_i' ($n_i' = n_i - \delta_i$)	p_i ($p_i = n_i'/n_i$)	$\hat{S}(y) = \prod_{i=1}^{k} p_i$
1	0.30	11	1	0	10	10/11	0.91
2	0.50	10	1	0	9	9/10	0.819
3	0.80	9	1	0	8	8/9	0.7272
4	0.90	8	2	0	6	6/8	0.5454
5	1.20	6	1	1	5	5/6	0.4545
6	1.34	4	1	0	3	3/4	0.341
7	1.40	3	1	0	2	2/3	0.227
8	1.50*	*	*	*	*	*	*

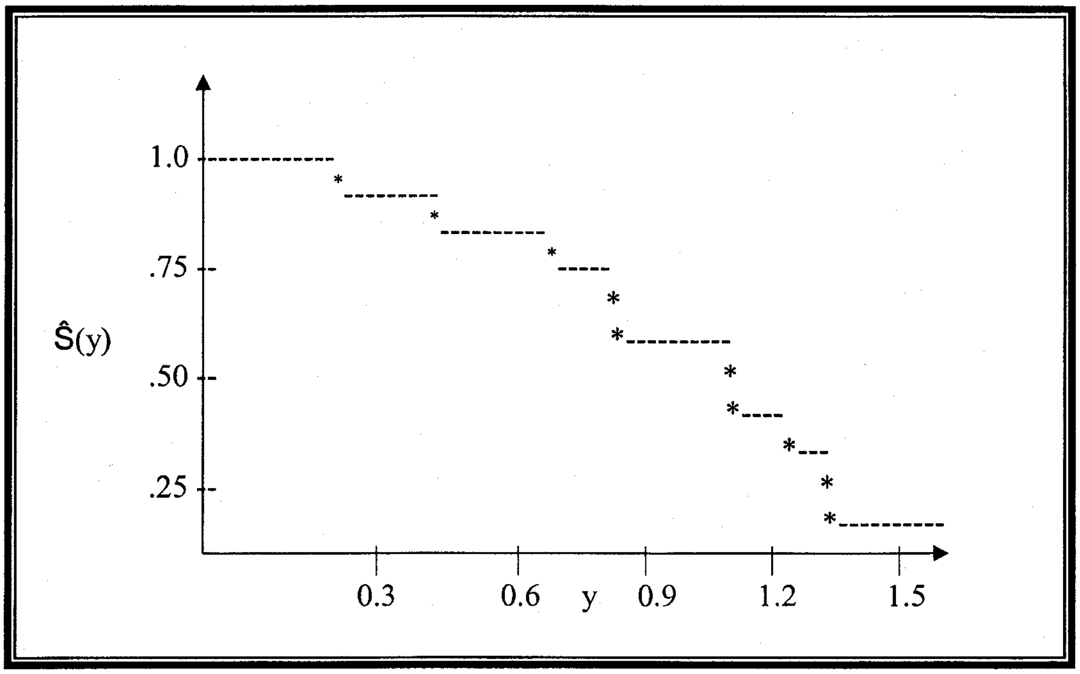


Figure 2.3.1 Estimated Survival Curve of Variable Y

The mean of Y is computed by calculating the area under the survival curve. Because the largest observation is censored, the mean can only be estimated up to $u_i = 1.40$ [4].

$$\mu_y = (1.00)(0.30) + (0.91)(0.50 - 0.30) + (0.819)(0.80 - 0.50) + (0.7272)(0.90 - 0.80)$$
$$+ (0.5454)(1.20 - 0.90) + (0.4545)(1.34 - 1.20) + (0.341)(1.40 - 1.34)$$
$$= 1.04813$$

The mean of X, which is of interest, is computed by taking the mean of Y (μ_y), and subtracting it from L = 1.60 derived in section 2.2.

$$\mu_x = L - \mu_y = 1.60 - 1.04813 = 0.55187$$

2.3.2 Computer Implementation

2.3.2.1 Minitab Implementation

K-M Estimates and Survival Plot can be computed by first inserting the observations of Y in one column and their censored values in another column. Select the following menus from the toolbar:

Stat → Reliability/Survival → Distribution Analysis (Right Censoring)

→ Nonparametric Distribution Analysis-Right Censoring

The K-M is the default Estimation Method, but the column containing the censored values must be indicated along with the value that defines the observation as right-censored (0). Figures 2.3.2.1(a) and 2.3.2.1(b) show the output after performing the above tasks.

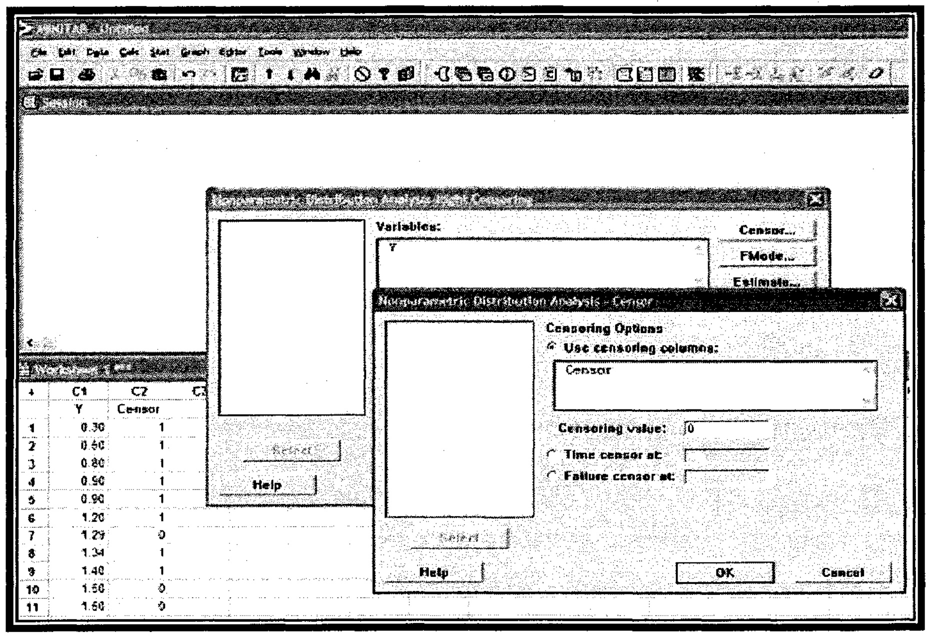


Figure 2.3.2.1(a) Using Minitab for K-M Estimates

Distribution Analysis: Y

Nonparametric Estimates

Characteristics of Variable Y

Mean(MTTF)	Standard Error	95.0% Normal CI Lower	Upper
1.04773	0.121396	0.809795	1.28566

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Prob.	Standard Error	95.0% Lower	Normal CI Upper
0.30	11	1	0.909091	0.086678	0.739204	1.00000
0.50	10	1	0.818182	0.116291	0.590255	1.00000
0.80	9	1	0.727273	0.134282	0.464086	0.99046
0.90	8	2	0.545455	0.150131	0.251202	0.83971
1.20	6	1	0.454545	0.150131	0.160293	0.74880
1.34	4	1	0.340909	0.149544	0.047809	0.63401
1.40	3	1	0.227273	0.136191	0.000000	0.49420

Figure 2.3.2.1(b) Minitab Output

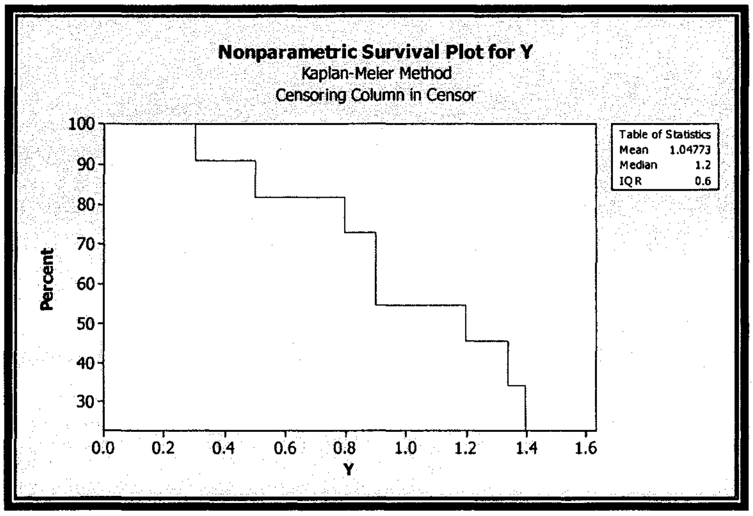


Figure 2.3.2.1(c) Minitab Survival Curve

2.3.2.2 SAS Implementaion

The LIFETEST procedure, given below, allows the use of the K-M method for right-censored data. The following statements in a SAS code perform the K-M method and output the survival estimates and the survival curve. The TIME statement is required and is used to define the variable Y and the value that indicates the observation is right-censored (0) [6]. Figures 2.3.2.2(a) and 2.3.2.2(b) show the SAS output after running the SAS code.

```
PROC LIFETEST DATA = <sas data set> METHOD = KM PLOT = (s);
TIME y*censor(0);
```

```
                         The LIFETEST Procedure
                    Product-Limit Survival Estimates

                            Standard       Number        Number
     y          Survival    Failure        Error         Failed       Left
  0.00000       1.0000         0              0            0          11
  0.30000       0.9091      0.0909        0.0867           1          10
  0.50000       0.8182      0.1818        0.1163           2           9
  0.80000       0.7273      0.2727        0.1343           3           8
  0.90000          .           .             .             4           7
  0.90000       0.5455      0.4545        0.1501           5           6
  1.20000       0.4545      0.5455        0.1501           6           5
  1.29000*         .           .             .             6           4
  1.34000       0.3409      0.6591        0.1495           7           3
  1.40000       0.2273      0.7727        0.1362           8           2
  1.50000*         .           .             .             8           1
  1.50000*         .           .             .             8           0
     NOTE: The marked survival times are censored observations.
                         Mean     Standard Error
                       1.04773      0.12140

NOTE: The mean survival time and its standard error were underestimated
    because the largest observation was censored and the estimation
          was restricted to the largest event time.
```

Figure 2.3.2.2(a) SAS Output

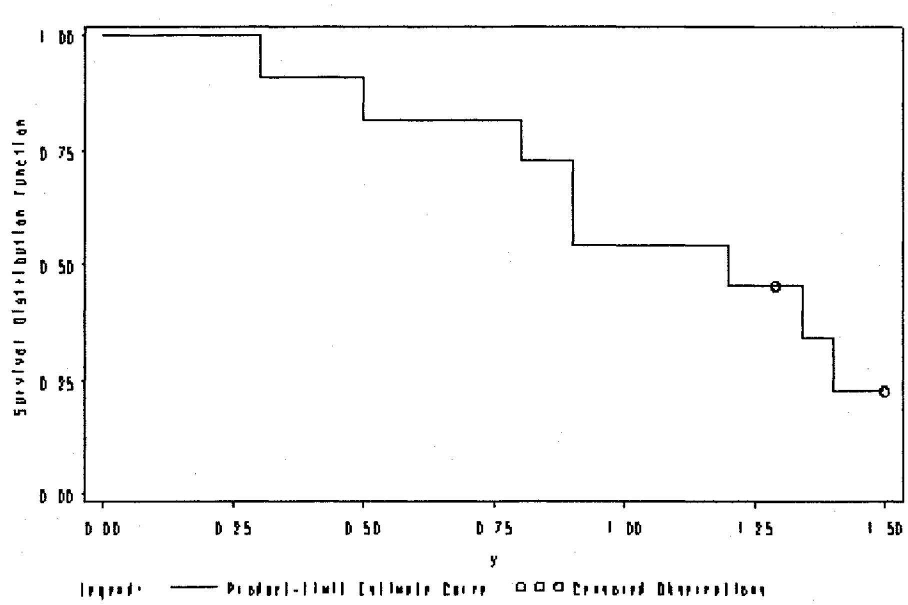


Figure 2.3.2.2(b) SAS Survival Curve

CHAPTER 3

MONTE CARLO SIMULATION EXPERIMENT

3.1 Monte Carlo Method

A Monte Carlo simulation experiment was developed to investigate the performance of the K-M method for computing a UCL of the population mean. The Monte Carlo method assures that if the input of a simulation is a random variable generated from a probability distribution and the simulation is repeated a large number of times, characteristics of the population will occur [5].

The steps in Monte Carlo Simulation experiment used in this thesis are described below:

1. Generate a pseudo-random sample of a specified sample size (N), from a specified probability distribution $f(x; \underline{\theta})$, where $\underline{\theta}$ represents the input vector of parameters. Select a value of DL, such that a specified percentage of nondetects (D) of the observations are less than DL. This will result in the sample $\{x_1, x_2, \ldots, x_N\}$ with D, which will be referred to as the input sample.
2. Generate a bootstrap sample $\{x_1^*, x_2^*, \ldots, x_N^*\}$ from the input sample.
3. Compute the K-M estimate of the survival function and also the area under the survival function S(t), which is an estimate of the population mean μ.
4. Repeat steps 2-3 a large number of times (B). This generates B estimates of the population mean ($\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_B$). Sort the B estimates of the population mean in

ascending order and extract the 95^{th} percentile. This is the 95% UCL.

5. Repeat steps 1-4 a large number of times (K). This generates K 95% UCLs. Compute the percentage of UCLs that are greater than the true mean. This generates the Estimated Coverage (%).

The above simulation experiment, programmed in SAS, was taking approximately two hours for one set of conditions. For this reason, we used B = 100 and K = 100.

3.2 Simulation Experiment

For the problem at hand we must know the true population parameters so that the performance of the K-M method can be investigated. For this reason, we simulate data from known distributions such as the ones seen in Figures 3.2.1(a), 3.2.1(b), and 3.2.1(c). We will generate a pseudo-random sample from one of these distributions of sample size N, ranging from 10 to 50 with percentage of D ranging from 10 to 50.

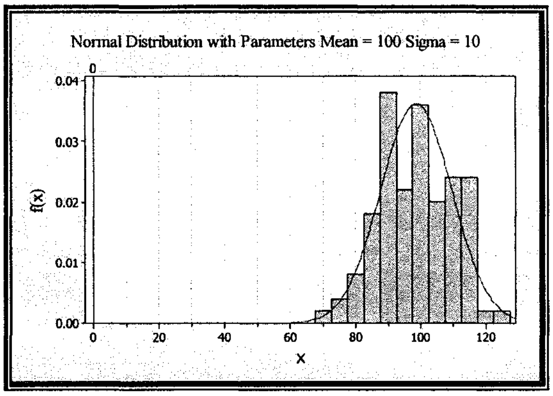


Figure 3.2(a) Histogram of a Sample Normal Distribution with Skewness (η) = 0 and Parameters μ = 100 σ = 10

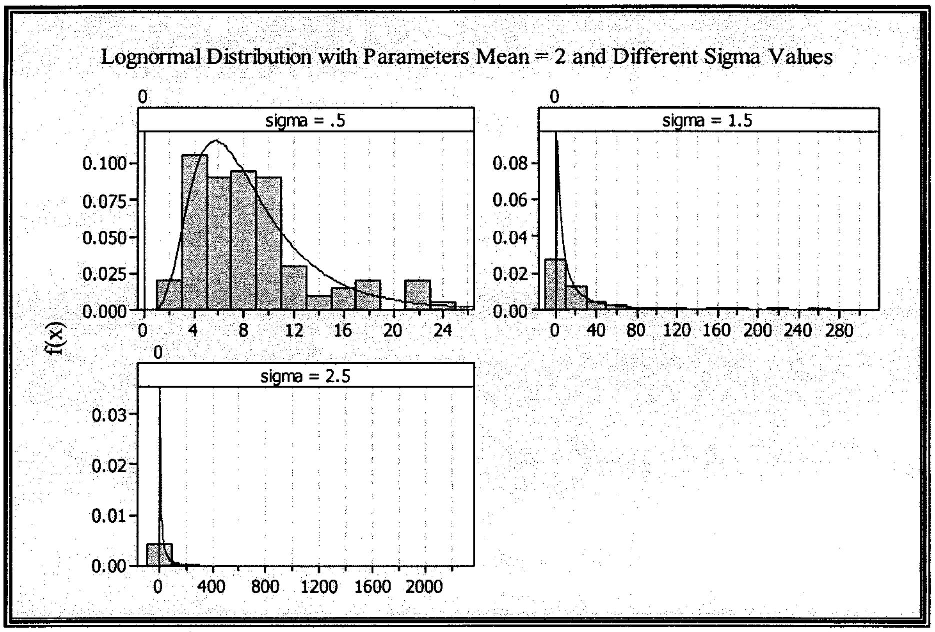


Figure 3.2(b) Histogram of a Sample Lognormal Distribution with Parameters μu = 2 and Different σ Values

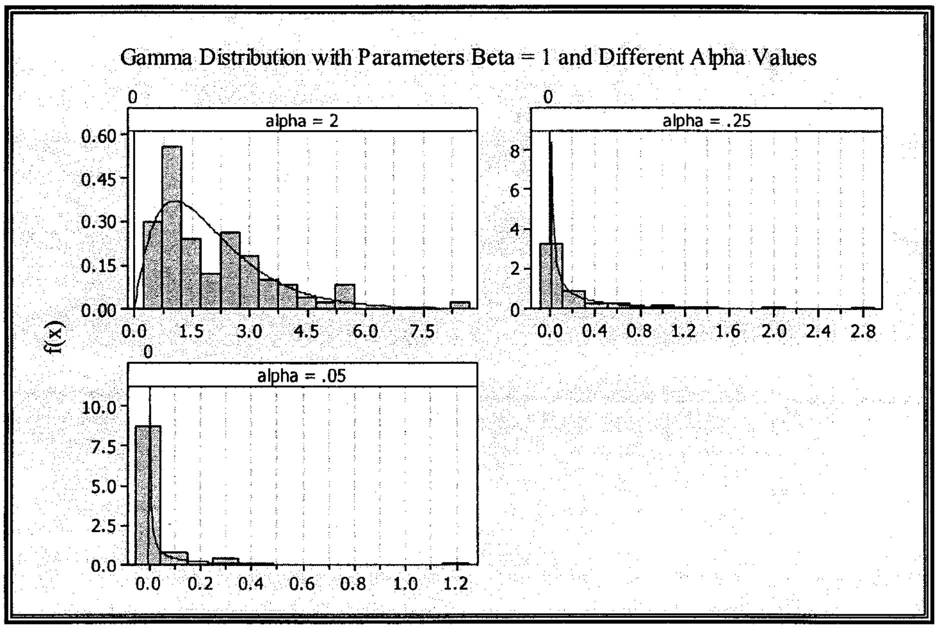


Figure 3.2(c) Histograms of Sample Gamma Distribution with Parameters β = 1 and Different α Values

In order to explain the steps of our simulation experiment for this thesis, we will use a data set generated from the Gamma distribution with skewness (η) = 1.41, parameters $\alpha = 2$ and $\beta = 1$, sample size (N) = 10, and nondetects (D) = 20%.

3.2.1 Monte Carlo Simulation Step 1

A data set generated from a Gamma distribution with parameters $\alpha = 2$ and $\beta = 1$ of N = 10, and sorted into ascending order by the variable x is shown below.

Obs	x
1	0.75026
2	0.97906
3	2.16379
4	2.25178
5	2.67793
6	2.79353
7	3.15135
8	4.03326
9	4.67053
10	4.96035

Figure 3.2.1(a) Computer-Generated Data Set

A left-censored environmental data set with D = 20% is of interest so the first two observations above are labeled as censored (DL = 1). The last eight observations are the uncensored observations. Figure 3.2.1(b) is the SAS computer-generated data set with the censor variable, whose value is 0 if the observation is left-censored and 1 if the observation is uncensored.

Obs	x	censor
1	<1	0
2	<1	0
3	2.16379	1
4	2.25178	1
5	2.67793	1
6	2.79353	1
7	3.15135	1
8	4.03326	1
9	4.67053	1
10	4.96035	1

Figure 3.2.1(b) Computer-Generated Data Set with Censor Variable

3.2.2 Monte Carlo Simulation Step 2

From the computer-generated data set shown in Figure 3.2.1(b), a boot sample of the same size is created by taking the nondetect observations and placing them as the nondetects for the boot sample. The uncensored observations of the boot sample are created by performing sampling with replacement from the uncensored observations of the computer-generated data set. This is accomplished by using a pseudo-random generator from a Uniform distribution, allowing the probability of an observation being chosen to be 1/8. It can be seen in Figure 3.2.2 that the nondetects of the boot sample are the same as those from the computer-generated data set.

Obs	x	censor
1	<1	0
2	<1	0
3	2.16379	1
4	2.16379	1
5	2.16379	1
6	2.67793	1
7	2.79353	1
8	4.03326	1
9	4.96035	1
10	4.96035	1

Figure 3.2.2 Boot Sample 1 Data Set

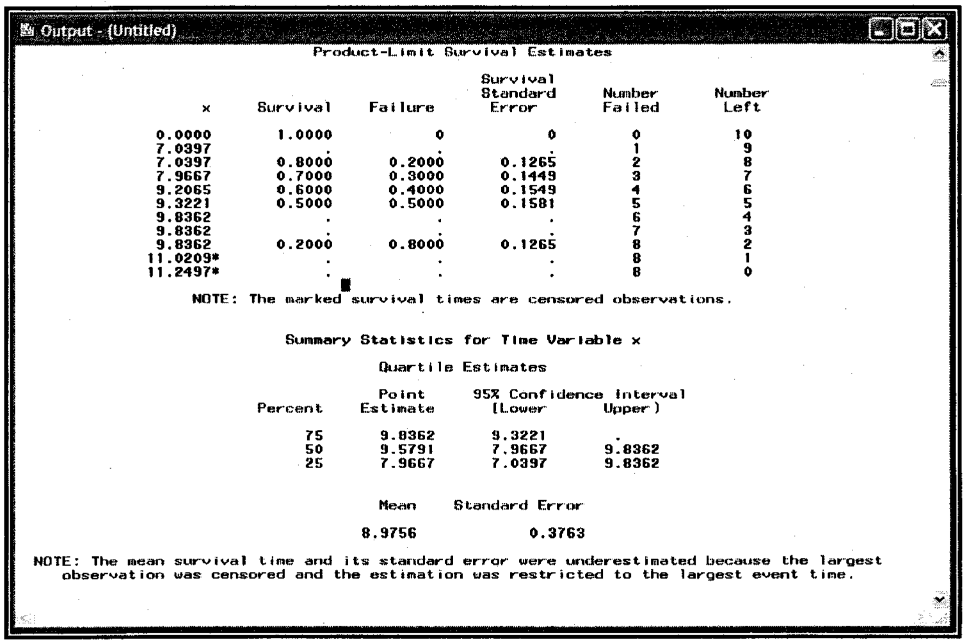
3.2.3 Monte Carlo Simulation Step 3

Before performing the K-M method on the left-censored boot sample, it will need to be transformed into a right-censored data set (Y). Using the technique discussed in Section 2.3, we see that the maximum observation is M = 4.96035. By letting $\varepsilon = 7.04$, L = M + ε = 12. The right-censored data set is created by subtracting each observation in the left-censored boot sample from L =12.

Obs	y	censor
1	7.0397	1
2	7.0397	1
3	7.9667	1
4	9.2065	1
5	9.3221	1
6	9.8362	1
7	9.8362	1
8	9.8362	1
9	>11	0
10	>11	0

Figure 3.2.3(a) Right-Censored Transformed Boot Sample

The K-M is performed on the right-censored data set (Y) as discussed in Section 2.3 to estimate the mean. The output of the SAS implemented K-M on the right-censored data set (Y) is seen in Figure 3.2.3(b) where the mean is computed to be 8.9756.



Output - (Untitled)

Product-Limit Survival Estimates

x	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	10
7.0397	.	.	.	1	9
7.0397	0.8000	0.2000	0.1265	2	8
7.9667	0.7000	0.3000	0.1449	3	7
9.2065	0.6000	0.4000	0.1549	4	6
9.3221	0.5000	0.5000	0.1581	5	5
9.8362	.	.	.	6	4
9.8362	.	.	.	7	3
9.8362	0.2000	0.8000	0.1265	8	2
11.0209*	.	.	.	8	1
11.2497*	.	.	.	8	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable x

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	9.8362	9.3221	.
50	9.5791	7.9667	9.8362
25	7.9667	7.0397	9.8362

Mean	Standard Error
8.9756	0.3763

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Figure 3.2.3(b) Output of the SAS Implemented K-M on the Right-Censored Data Set (Y)

The mean of the boot sample is computed by taking the mean of Y and subtracting it from L.

$$\mu_x = L - \mu_y = 12 - 8.9756 = 3.0244$$

3.2.4 Monte Carlo Simulation Step 4

Repeat Sections 3.2.2 through 3.2.3 100 times. This generates 100 estimates of the population mean. After sorting the 100 means in ascending order, the 95th percentile is the 95% UCL of the mean in the presence of nondetects in environmental data set generated in Section 3.2.1.

3.2.5 Monte Carlo Simulation Step 5

Repeat Sections 3.2.1 through 3.2.4 100 times. This generates 100 95% UCLs. Each 95% UCL is tested to see if it is greater than the true mean, which in this case the true mean is,

$$\mu = \alpha\beta = 2 .$$

The Estimated Coverage is the percentage of the 100 95% UCLs that are greater than μ. Figure 3.2.5 is a flowchart of the Monte Carlo Simulation Experiment.

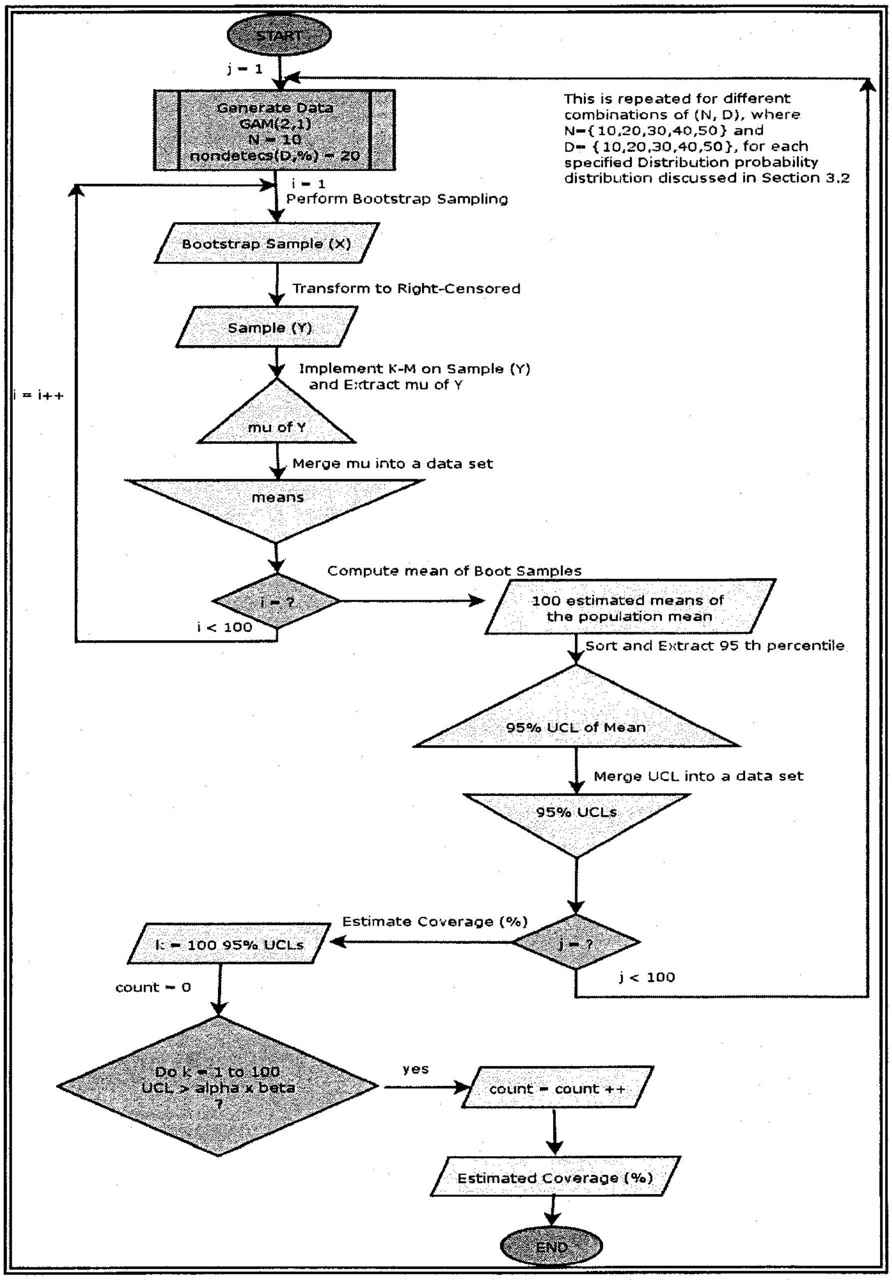


Figure 3.2.5 Flow Chart of Monte Carlo Simulation Experiment

Chapter 4

RESULTS

For each distribution in combination with the different sample sizes, and different D% nondetects, the following will be presented: a table of the summary statistics of the bootstrap UCLs for the mean; a table of the Estimated Coverage (%) as a function of Nondetects (%); a graph of the Estimated Coverage (%) vs. D (%) grouped by Sample Size. For each case, the distribution's skewness (η) will be observed to see its effect on the accuracy of the K-M method for computing the UCL for the population mean from environmental samples with nondetects.

4.1 Normal Distribution with $\eta = 0$, $\mu = 100$, and $\sigma = 10$

In this section, input samples are generated from a Normal distribution with parameters $\mu = 100$ and $\sigma = 10$, and the K-M method combined with bootstrap as explained in detail in Chapter 3 is used. The results are summarized in Tables 4.1(a) and 4.1(b). It can be seen from Table 4.1(a) that the mean UCL exceeds the true mean of 100 by no more than 9%. It is seen from Table 4.1(b) that when the underlying distribution is normal with $\eta = 0$, the K-M method generally gives coverage greater than or equal to the specified confidence (95%). Figure 4.1 is a graph of the estimated coverage probabilities shown in Table 4.1(b).

Table 4.1(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated Normal Distribution with $\eta = 0$, $\mu = 100$, and $\sigma = 10$

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	105.78	0.313	3.13	98.19	114.98
10	20	103.52	0.207	2.07	96.92	109.82
10	30	103.08	0.187	1.87	98.79	108.76
10	40	102.27	0.164	1.64	98.10	105.93
10	50	102.25	0.150	1.50	97.58	105.93
20	10	105.50	0.354	3.54	98.31	114.91
20	20	103.97	0.239	2.39	97.40	109.72
20	30	103.58	0.198	1.98	97.92	108.96
20	40	103.33	0.181	1.81	99.86	107.39
20	50	102.82	0.126	1.26	100.02	106.72
30	10	105.61	0.334	3.34	98.86	113.12
30	20	104.23	0.208	2.08	97.78	109.37
30	30	104.08	0.190	1.90	98.27	108.44
30	40	103.38	0.160	1.60	98.74	107.83
30	50	103.71	0.155	1.55	100.50	107.50
40	10	107.33	0.294	2.94	100.63	113.93
40	20	104.94	0.250	2.50	98.57	111.58
40	30	104.89	0.204	2.04	99.20	109.78
40	40	104.83	0.176	1.76	100.73	109.56
40	50	103.86	0.169	1.69	99.96	108.62
50	10	108.26	0.410	4.10	96.56	120.16
50	20	107.01	0.265	2.65	100.21	114.90
50	30	106.20	0.224	2.24	101.25	111.72
50	40	105.75	0.173	1.73	102.13	111.06
50	50	105.40	0.156	1.56	101.80	109.94

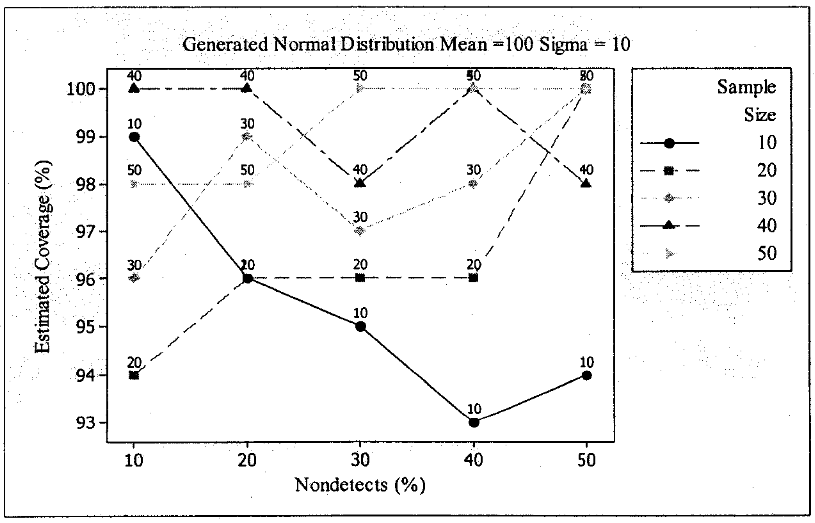


Figure 4.1 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.1(b) Estimated Coverage as a Function of Nondetects (%) for Generated Normal Distribution with $\eta = 0$

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	99
10	20	96
10	30	95
10	40	93
10	50	94
20	10	94
20	20	96
20	30	96
20	40	96
20	50	100
30	10	96
30	20	99
30	30	97
30	40	98
30	50	100
40	10	100
40	20	100
40	30	98
40	40	100
40	50	98
50	10	98
50	20	98
50	30	100
50	40	100
50	50	100

4.2 Lognormal Distribution

4.2.1 LN(2, 2.5) with $\eta = 11{,}824$ and $\mu = 168.174$

In this section, input samples are generated from a Lognormal distribution with parameters $\mu u = 2$ and $\sigma = 2.5$. The results are summarized in Tables 4.2.1(a) and 4.2.1(b). It can be seen from Table 4.2.1(b) that when the underlying distribution is Lognormal with parameters $\mu u = 2$ and $\sigma = 2.5$, the K-M method gives coverage a lot smaller than the specified confidence (95%). This is due to the fact that the Lognormal distribution with parameters $\mu u = 2$ and $\sigma = 2.5$ is heavily skewed $\eta = 11{,}824$. Figure 4.2.1 is a graph of the estimated coverage probabilities shown in Table 4.2.1(b)

Table 4.2.1(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated LN(2, 2.5) with η = 11,824 and μ = 168.174

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	478	179	1792	3.44	16400
10	20	397	140	1397	9.19	13705
10	30	273.3	57.1	571.2	4.57	3294.1
10	40	337.6	99.5	995.4	7.22	8409.8
10	50	638	419	4190	7.86	42018
20	10	293.3	42.7	426.6	14.1	2463.0
20	20	426.5	90.0	899.5	13.7	6049.1
20	30	412	111	1106	21.5	8271
20	40	308.9	65.6	655.6	5.23	5897.5
20	50	305.2	43.0	430.3	9.15	2416.0
30	10	305.6	41.1	410.7	32.8	2691.3
30	20	357.0	78.2	782.2	22.8	7157.7
30	30	312.7	72.2	721.6	12.4	6191.7
30	40	323.2	53.2	531.7	19.3	3702.1
30	50	292.1	44.9	448.6	14.4	3553.4
40	10	400.1	92.4	924.4	41.7	6790.3
40	20	326.8	51.8	518.4	27.1	4196.4
40	30	467	209	2086	27.5	20692
40	40	379.8	77.6	775.9	13.9	6490.9
40	50	385.3	54.8	548.4	38.9	3204.0
50	10	297.1	35.9	358.6	30.0	2623.9
50	20	250.8	27.8	278.0	26.1	1590.6
50	30	305.5	40.2	402.1	27.9	2770.5
50	40	320.0	48.1	481.0	23.3	3132.2
50	50	266.0	36.0	360.1	26.4	2718.0

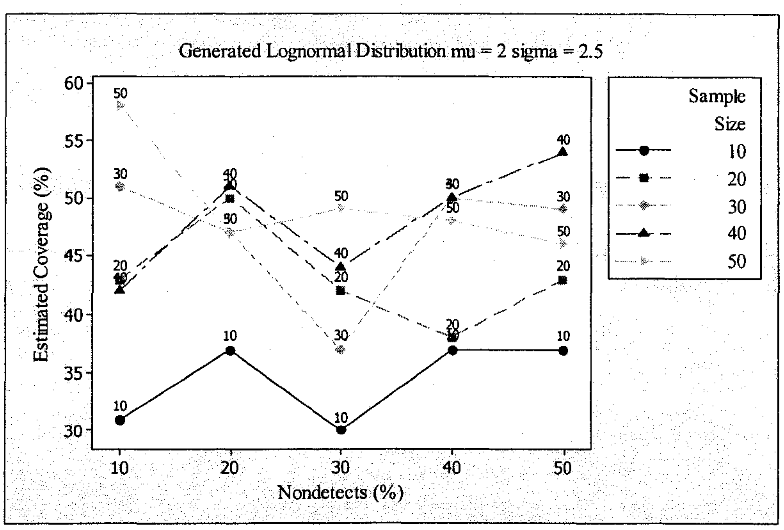


Figure 4.2.1 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.2.1(b) Estimated Coverage as a Function of Nondetects (%) for Generated Lognormal Distribution with η = 11,824

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	31
10	20	37
10	30	30
10	40	37
10	50	37
20	10	43
20	20	50
20	30	42
20	40	38
20	50	43
30	10	51
30	20	47
30	30	37
30	40	50
30	50	49
40	10	42
40	20	51
40	30	44
40	40	50
40	50	54
50	10	58
50	20	47
50	30	49
50	40	48
50	50	46

4.2.2 LN(2, 1.5) with η = 33.468 and μ = 22.7599

In this section, input samples are generated from a Lognormal distribution with parameters μu = 2 and σ = 1.5, and the K-M method combined with the bootstrap method is used. It is seen from Table 4.2.2(b) that when the underlying distribution is Lognormal with parameters μu = 2 and σ = 1.5, the K-M method gives coverage that improves when compared to the previous case in Section 4.2.1 but is still smaller than the specified confidence (95%). This is due to the fact that the Lognormal distribution with parameters μu = 2 and σ = 1.5 is skewed η = 33.468 but less skewed than the case in Section 4.2.1. Figure 4.2.2 is a graph of the estimated coverage probabilities shown in Table 4.2.2(b).

Table 4.2.2(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated LN(2, 1.5) with η = 33.368 and μ = 22.7599

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	48.00	7.52	75.17	7.30	683.28
10	20	38.68	3.18	31.79	7.52	162.24
10	30	44.17	8.28	82.82	6.06	782.05
10	40	44.17	8.28	82.82	6.06	782.05
10	50	56.17	5.91	59.11	5.49	371.74
20	10	40.43	2.73	27.31	12.00	153.15
20	20	38.28	2.36	23.56	8.27	121.16
20	30	43.48	3.49	34.86	11.62	193.66
20	40	38.15	2.10	20.99	6.75	116.30
20	50	40.21	2.47	24.65	8.55	168.73
30	10	34.97	2.07	20.65	12.59	168.03
30	20	42.52	3.26	32.63	14.96	211.00
30	30	35.64	2.10	21.04	12.45	159.94
30	40	36.15	1.88	18.83	12.20	104.07
30	50	34.32	1.90	19.01	11.04	122.33
40	10	37.68	2.26	22.56	11.15	142.89
40	20	34.34	1.40	14.03	13.28	89.08
40	30	34.26	1.52	15.21	14.15	83.75
40	40	37.43	2.23	22.32	11.24	147.36
40	50	40.80	3.09	30.87	12.42	236.77
50	10	33.03	2.13	21.27	12.00	181.07
50	20	33.97	1.53	15.27	12.41	83.53
50	30	33.15	1.52	15.16	17.55	112.95
50	40	39.98	3.97	39.71	14.69	404.99
50	50	35.11	2.02	20.20	12.18	192.33

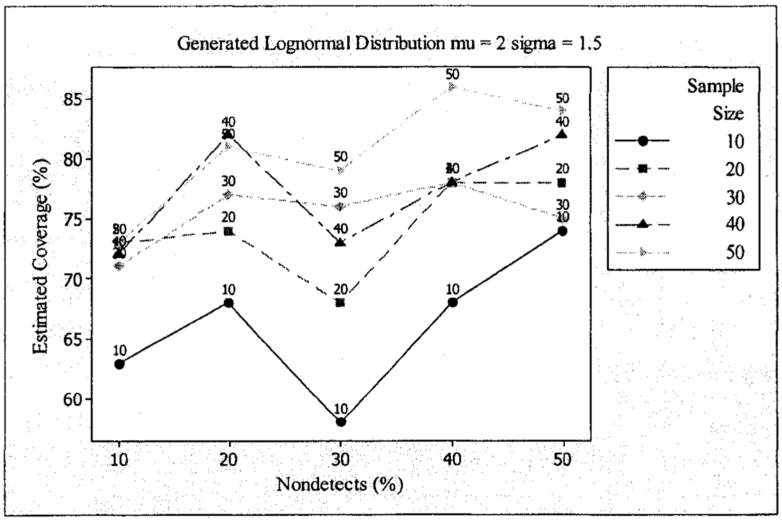


Figure 4.2.2 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.2.2(b) Estimated Coverage as a Function of Nondetects (%) for Generated Lognormal Distribution with η = 33.368

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	63
10	20	68
10	30	58
10	40	68
10	50	74
20	10	73
20	20	74
20	30	68
20	40	78
20	50	78
30	10	71
30	20	77
30	30	76
30	40	78
30	50	75
40	10	72
40	20	82
40	30	73
40	40	78
40	50	82
50	10	73
50	20	81
50	30	79
50	40	86
50	50	84

4.2.3 LN(2, 0.5) with η = 1.75 and μ = 8.3729

Input samples are generated from a Lognormal distribution with parameters μu = 2 and σ = 0.5, and the K-M method combined with the bootstrap method is used. The results are summarized in Tables 4.2.3(a) and 4.2.3(b). It can be seen from Table 4.2.3(b) that when the underlying distribution is Lognormal with parameters μu = 2 and σ = 0.5, the K-M method gives coverage that improves when compared to the previous cases in Sections 4.2.1 and 4.2.2. The coverage comes very close to the specified confidence (95%). This is due to the fact that the Lognormal distribution with μu = 2 and σ = 0.5 is somewhat symmetric but still contains a small positive skewness of η =1.75. Figure 4.2.3 is a graph of the estimated coverage probabilities shown in Table 4.2.3(b).

Table 4.2.3(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated LN(2, 0.5) with η = 1.75 and μ = 8.3729

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	10.051	0.187	1.873	5.459	16.817
10	20	10.508	0.200	1.995	6.987	19.372
10	30	10.627	0.210	2.097	6.680	16.531
10	40	11.396	0.253	2.526	6.686	20.173
10	50	11.988	0.245	2.449	7.806	19.068
20	10	10.011	0.126	1.260	7.615	14.745
20	20	10.317	0.147	1.472	7.105	14.633
20	30	10.267	0.136	1.358	7.554	14.006
20	40	10.375	0.138	1.377	6.924	14.638
20	50	11.046	0.158	1.583	6.852	16.178
30	10	9.745	0.106	1.062	7.188	12.360
30	20	9.602	0.101	1.009	7.137	12.091
30	30	9.785	0.104	1.036	7.861	12.878
30	40	9.936	0.103	1.026	7.509	13.269
30	50	10.354	0.115	1.153	7.587	14.386
40	10	9.6291	0.080	0.8004	7.585	11.683
40	20	9.4913	0.0791	0.7913	7.7725	11.975
40	30	9.811	0.101	1.006	7.253	12.966
40	40	10.147	0.103	1.034	7.858	13.157
40	50	10.426	0.0964	0.964	8.613	13.435
50	10	9.3529	0.0820	0.8203	7.8980	11.429
50	20	9.5451	0.0792	0.7919	7.9804	11.737
50	30	9.7071	0.0847	0.8475	7.5751	12.373
50	40	9.8876	0.0931	0.9311	8.1642	13.156
50	50	10.206	0.0726	0.726	8.555	12.569

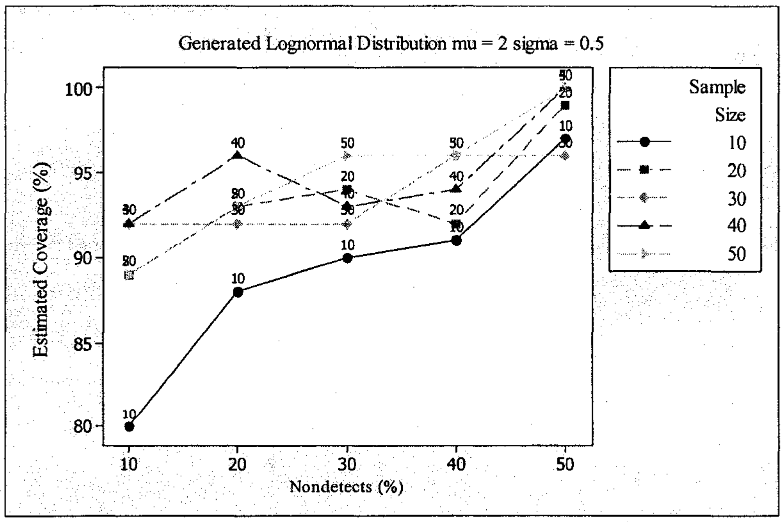


Figure 4.2.3 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.2.3(b) Estimated Coverage as a Function of Nondetects (%) for Generate Lognormal Distribution with $\eta = 1.75$

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	80
10	20	88
10	30	90
10	40	91
10	50	97
20	10	89
20	20	93
20	30	94
20	40	92
20	50	99
30	10	92
30	20	92
30	30	92
30	40	96
30	50	96
40	10	92
40	20	96
40	30	93
40	40	94
40	50	100
50	10	89
50	20	93
50	30	96
50	40	96
50	50	100

4.3 Gamma Distribution

4.3.1 GAM(.05, 1) with $\eta = 8.944$ and $\mu = .05$

In this section, input samples are generated data set from a Gamma distribution with parameters $\alpha = 0.05$ and $\beta = 1$, and the K-M method combined with the bootstrap method is used. The results are summarized in Tables 4.3.1(a) and 4.3.1(b). It is seen from Table 4.3.1(b) that when the underlying distribution is Gamma with parameters $\alpha = 0.05$ and $\beta = 1$, the K-M method gives coverage a lot smaller than the specified confidence (95%). This is due to the fact that the Gamma distribution with parameters $\alpha = 0.05$ and $\beta = 1$ is quite skewed with $\eta = 8.944$. Figure 4.3.1 is a graph of the estimated coverage probabilities shown in Table 4.3.1(b).

Table 4.3.1(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated GAM(.05, 1) with η = 8.944 and μ = .05

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	0.1344	0.0357	0.2082	0.000	1.0530
10	20	0.0941	0.0203	0.1183	0.000	0.4790
10	30	0.1043	0.0220	0.1284	0.000	0.4750
10	40	0.1298	0.0292	0.1701	0.000	0.6900
10	50	0.0953	0.0201	0.1170	0.000	0.4840
20	10	0.0885	0.0168	0.0978	0.000	0.4810
20	20	0.1210	0.0185	0.1076	0.008	0.4650
20	30	0.0870	0.0130	0.0759	0.000	0.2740
20	40	0.0862	0.0147	0.0857	0.007	0.3300
20	50	0.0895	0.0154	0.0901	0.001	0.3730
30	10	0.1016	0.0139	0.0810	0.006	0.3690
30	20	0.0921	0.0137	0.0800	0.002	0.3570
30	30	0.1234	0.0152	0.0886	0.009	0.3380
30	40	0.0974	0.0132	0.0767	0.004	0.3490
30	50	0.0947	0.0133	0.0776	0.007	0.3270
40	10	0.08576	0.00832	0.04849	0.010	0.2050
40	20	0.1036	0.0111	0.0649	0.003	0.2640
40	30	0.0848	0.0102	0.0595	0.013	0.2230
40	40	0.1063	0.0123	0.0715	0.012	0.3040
40	50	0.1149	0.0161	0.0938	0.006	0.4530
50	10	0.07935	0.00693	0.04039	0.018	0.2280
50	20	0.09079	0.00931	0.05427	0.010	0.2190
50	30	0.0888	0.0108	0.0627	0.010	0.2380
50	40	0.09091	0.00991	0.05778	0.002	0.2600
50	50	0.1062	0.0107	0.0626	0.019	0.2840

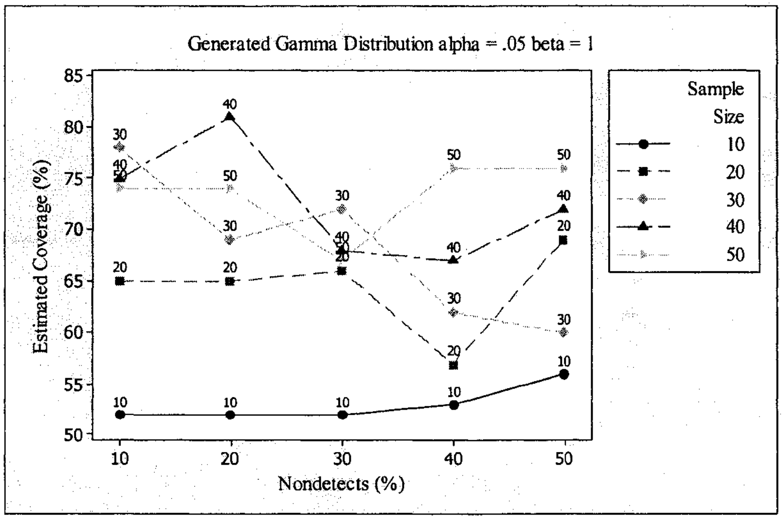


Figure 4.3.1 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.3.1(b) Estimated Coverage as a Function of Nondetects (%) for Generated Gamma Distribution with η = 8.944

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	52
10	20	52
10	30	52
10	40	53
10	50	56
20	10	65
20	20	65
20	30	66
20	40	57
20	50	69
30	10	78
30	20	69
30	30	72
30	40	62
30	50	60
40	10	75
40	20	81
40	30	68
40	40	67
40	50	72
50	10	74
50	20	74
50	30	67
50	40	76
50	50	76

4.3.2 GAM(0.25, 1) with $\eta = 4$ and $\mu = 0.25$

In this section, input samples are generated from a Gamma distribution with parameters $\alpha = 0.25$ and $\beta = 1$, and the K-M method combined with bootstrap is used. The results are summarized in Tables 4.3.2(a) and 4.3.2(b). It can be seen from Table 4.3.2(b) that when the underlying distribution is Gamma with parameters $\alpha = 0.25$ and $\beta = 1$, the K-M method gives coverage that improves when compared to the previous case in Section 4.3.1 but is still smaller than the specified confidence (95%). This is due to the fact that the Gamma distribution with parameters $\alpha = 0.25$ and $\beta = 1$ is skewed $\eta = 4$, but less skewed than the case in Section 4.3.1. Figure 4.3.2 is a graph of the estimated coverage probabilities shown in Table 4.3.2(b).

Table 4.3.2(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated GAM(0.25, 1) with $\eta = 4$ and $\mu = 0.25$

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	0.3835	0.0263	0.2633	0.0400	1.427
10	20	0.4541	0.0328	0.3277	0.0250	1.737
10	30	0.4338	0.0258	0.2580	0.0780	1.455
10	40	0.4479	0.0280	0.2800	0.0140	1.324
10	50	0.5471	0.0351	0.3510	0.0760	1.850
20	10	0.4271	0.0207	0.2066	0.0580	1.024
20	20	0.4033	0.0175	0.1751	0.0980	0.931
20	30	0.4629	0.0226	0.2260	0.0290	1.398
20	40	0.4158	0.0209	0.2093	0.1040	1.152
20	50	0.4556	0.0257	0.2573	0.1090	1.789
30	10	0.3825	0.0140	0.1395	0.1420	0.715
30	20	0.3874	0.0131	0.1307	0.1550	0.701
30	30	0.3923	0.0171	0.1711	0.1280	1.002
30	40	0.4013	0.0163	0.1631	0.1310	0.847
30	50	0.4117	0.0150	0.1503	0.1700	0.883
40	10	0.3640	0.0122	0.1222	0.1050	0.641
40	20	0.3797	0.0128	0.1276	0.1270	0.718
40	30	0.3536	0.0116	0.1156	0.1380	0.798
40	40	0.3766	0.0121	0.1214	0.1510	0.752
40	50	0.3860	0.0123	0.1230	0.1560	0.823
50	10	0.34558	0.00939	0.09393	0.17500	0.676
50	20	0.35490	0.00918	0.09178	0.1810	0.586
50	30	0.3659	0.0119	0.1186	0.1640	0.796
50	40	0.36018	0.00864	0.08642	0.18500	0.677
50	50	0.3791	0.0115	0.1149	0.1720	0.735

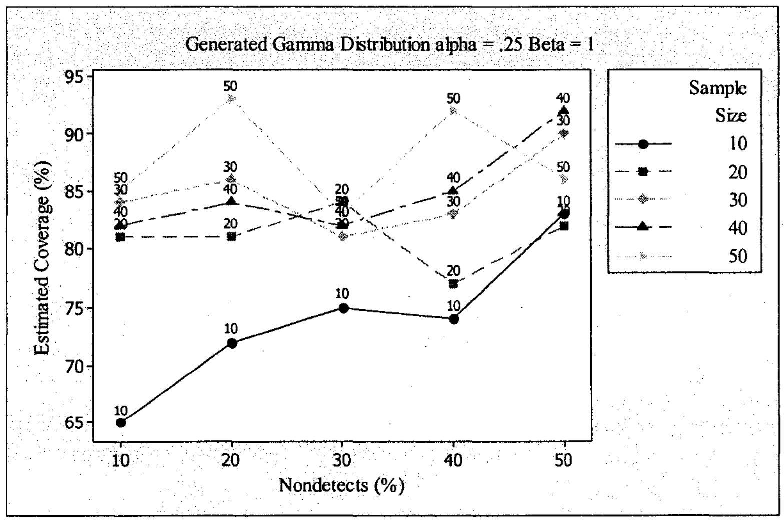


Figure 4.3.2 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.3.2(b) Estimated Coverage as a Function of Nondetects (%) for Generated Gamma Distribution with $\eta = 4$

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	65
10	20	72
10	30	75
10	40	74
10	50	83
20	10	81
20	20	81
20	30	84
20	40	77
20	50	82
30	10	84
30	20	86
30	30	81
30	40	83
30	50	90
40	10	82
40	20	84
40	30	82
40	40	85
40	50	92
50	10	85
50	20	93
50	30	83
50	40	92
50	50	86

4.3.3 GAM(2, 1) with $\eta = 1.414$ and $\mu = 2$

Input samples are generated from a Gamma distribution with parameters $\alpha = 2$ and $\beta = 1$, and the K-M method combined with the bootstrap method is used. The results are summarized in Tables 4.3.3(a) and 4.3.3(b). It is seen from Table 4.3.3(b) that when the underlying distribution is Gamma with parameters $\alpha = 2$ and $\beta = 1$, the K-M method gives coverage that improves when compared to the previous cases in Sections 4.3.1 and 4.3.2. The coverage comes very close to the specified confidence (95%). This is due to the fact that the Gamma distribution with parameters $\alpha = 2$ and $\beta = 1$ is more symmetric than the other gamma cases, but still has a positive skewness of $\eta = 1.414$. Figure 4.2.3 is a graph of the estimated coverage probabilities shown in Table 4.3.3(b).

Table 4.3.3(a) Summary Statistics of Bootstrap UCLs of the Mean from a Generated GAM(2, 1) with $\eta = 1.414$ and $\mu = 2$

Sample Size	Nondetects (%)	Mean	SE Mean	StDev	Min	Max
10	10	2.6856	0.0652	0.6516	1.1590	4.2880
10	20	2.5794	0.0617	0.6173	1.3370	4.2120
10	30	2.8561	0.0696	0.6960	1.3600	4.6540
10	40	2.8071	0.0697	0.6971	1.3170	4.9470
10	50	3.1074	0.0826	0.8259	1.4750	5.8610
20	10	2.4515	0.0411	0.4112	1.5550	3.6570
20	20	2.5122	0.0426	0.4263	1.6270	3.5640
20	30	2.6237	0.0444	0.4436	1.5520	3.8110
20	40	2.6433	0.0460	0.4597	1.6930	4.0400
20	50	2.8359	0.0480	0.4798	1.7440	4.6540
30	10	2.4456	0.0363	0.3628	1.6560	3.4330
30	20	2.4386	0.0296	0.2959	1.7520	3.6140
30	30	2.4761	0.0302	0.3017	1.7910	3.3650
30	40	2.5767	0.0351	0.3512	1.7750	3.3990
30	50	2.6457	0.0321	0.3212	1.7630	3.3690
40	10	2.3470	0.0280	0.2798	1.5480	2.9910
40	20	2.3410	0.0267	0.2670	1.7810	3.0070
40	30	2.4252	0.0295	0.2953	1.8490	3.3320
40	40	2.5647	0.0297	0.2966	1.9460	3.3280
40	50	2.7015	0.0295	0.2952	1.9680	3.5650
50	10	2.3358	0.0242	0.2416	1.7040	3.0200
50	20	2.3554	0.0253	0.2526	1.7570	3.0840
50	30	2.3855	0.0231	0.2306	1.7790	2.9830
50	40	2.4879	0.0247	0.2466	1.8810	3.2130
50	50	2.6579	0.0296	0.2957	1.8990	3.4440

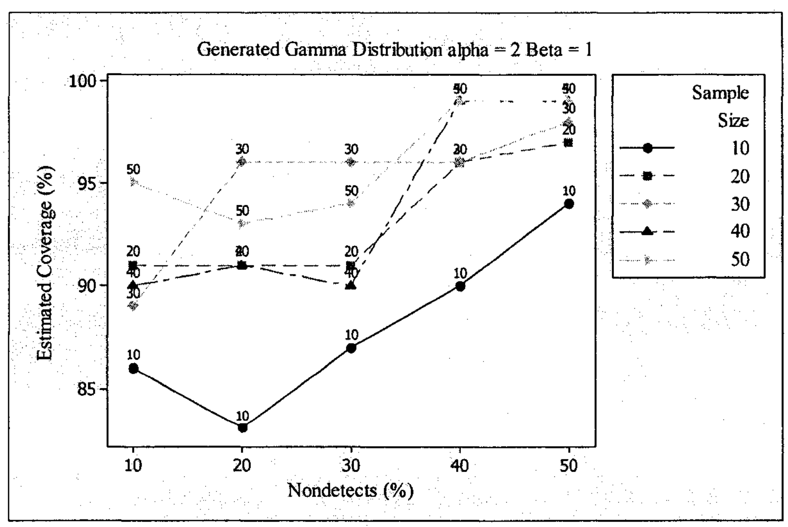


Figure 4.3.3 Scatterplot of Estimated Coverage (%) vs Nondetects (%) for Different Sample Size

Table 4.3.3(b) Estimated Coverage as a Function of Nondetects (%) for Generated Gamma Distribution with $\eta = 1.414$

Sample Size	Nondetects (%)	Estimated Coverage (%)
10	10	86
10	20	83
10	30	87
10	40	90
10	50	94
20	10	91
20	20	91
20	30	91
20	40	96
20	50	97
30	10	89
30	20	96
30	30	96
30	40	96
30	50	98
40	10	90
40	20	91
40	30	90
40	40	99
40	50	99
50	10	95
50	20	93
50	30	94
50	40	99
50	50	99

4.4 A Look At How Skewness Affects Estimated Coverage

After looking at the results from the simulation experiment, I began to see a strong relation between the coverage and the skewness of the input sample's probability distribution. When the input sample is generated from a probability distribution that is symmetric, the coverage is more or less at the specifed 95%. By generating a sample input from a proability distribution that has a positive skewness, the coverage begins to decrease. The higher the skewness the poorer the coverage. Table 4.4 shows the Estimated Coverage for the different probability distributions with there respective skewness that the input samples were generated from. The order is in increasing skewness from left to right.

Table 4.4 A Look At How Skewness Affects Estimated Coverage

Sample Size N	Nondetects D (%)	N(100,10)	GAM(2, 1)	LN(2, 0.5)	GAM(.25, 1)	GAM(.05, 1)	LN(2, 1.5)	LN(2, 2.5)
		$\eta = 0$	$\eta = 1.414$	$\eta = 1.75$	$\eta = 4$	$\eta = 8.944$	$\eta = 33.468$	$\eta = 11{,}824$
10	10	99	86	80	65	52	63	31
10	20	96	83	88	72	52	68	37
10	30	95	87	90	75	52	58	30
10	40	93	90	91	74	53	68	37
10	50	94	94	97	83	56	74	37
20	10	94	91	89	81	65	73	43
20	20	96	91	93	81	65	74	50
20	30	96	91	94	84	66	68	42
20	40	96	96	92	77	57	78	38
20	50	100	97	99	82	69	78	43
30	10	96	89	92	84	78	71	51
30	20	99	96	92	86	69	77	47
30	30	97	96	92	81	72	76	37
30	40	98	96	96	83	62	78	50
30	50	100	98	96	90	60	75	49
40	10	100	90	92	82	75	72	42
40	20	100	91	96	84	81	82	51
40	30	98	90	93	82	68	73	44
40	40	100	99	94	85	67	78	50
40	50	98	99	100	92	72	82	54
50	10	98	95	89	85	74	73	58
50	20	98	93	93	93	74	81	47
50	30	100	94	96	83	67	79	49
50	40	100	99	96	92	76	86	48
50	50	100	99	100	86	76	84	46

CHAPTER 5

CONCLUSION

There are a number of methods implemented for interpreting and analyzing environmental data that fall below the DL (nondetects). The following two are the most common methods: ignoring the observations that fall below the DL and the substitution method.

It has been proposed in many environmental articles as well as the recently published book, Nondetects and Data Analysis: Statistics for Censored Environmental Data by Dennis R. Helsel, to use the K-M method for estimating summary statistics on environmental samples. The book recommends the use of the K-M for any sized left-censored environmental data sets as long as the percentage of nondetects is less than 50%.

The simulation experiments conducted in this thesis show that the K-M method works well as long as the population distribution is symmetric. After simulating an input sample from a normal distribution, three different skewed lognormal distributions, and three different skewed gamma distributions with sample sizes ranging from N = 10 to N =50 and nondetects ranging from D = 10% to D = 50% the simulation experiment shows the following:

1. The K-M method combined with bootstrap, can be used to get a confidence interval for the mean of a population.

2. The coverage of this confidence is at least as large as the specified coverage when the data distribution is symmetric (skewness = 0).

3. The coverage begins to decrease as skewness increases.

4. The coverage is much lower than the specified confidence (95%) when skewness is high.

Therefore, in order to use the K-M method on left-censored environmental data, not only does the data set need to be transformed to a right-censored data set, but it must come from a population whose probability distribution is somewhat symmetric.

This experiment was unable to be evaluated analytically. The problem with estimating the mean of a data set that is heavily skewed with a high percentage of accuracy can be seen from the graph below. Below is a graph of a lognormal distribution with $\eta = 11{,}824$ and $\mu = 168$. The mean is found on the right-hand tail of the distribution making it near impossible to accurately estimate it analytically. Considering that many of the environmental data sets are heavily skewed, estimating the mean is still a problem for environmental scientists.

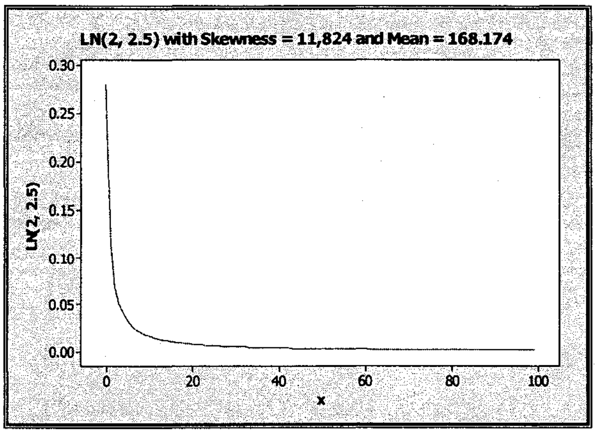


Figure 5 Estimated Distribution of LN(2, 2.5) with $\eta = 11{,}824$ and $\mu = 168.174$

APPENDIX I

SAS SOURCE CODE

A copy of the SAS code used to obtain the results in Chapter 4 are presented below. Due to limited resources, the program was edited and run for each distribution discussed in Chapter 3.

```
/*******Generated Gamma Distribution w/ beta = 1 alpha = 2*******/

%LET beta = 1;
%LET alpha = 2;

/****Boottest Macro****/

%macro boottest(size=, percent_Of_Detects=);

  data get_value;
    num = &size*&percent_Of_Detects;
    numb = .01*num;
    CALL SYMPUT('number',numb);
  run;

  /***100 Boot Samples Loop***/

  %DO i = 1 %to 100;

   /****Sampling w/ Replacement From Generated Data Set****/

   data boot;
     set computer_generated (FIRSTOBS = 1 OBS = &number);
     censor = 0;
   run;

   data boota;
     %do j = 1 %to (&size-&number);
       k = int(ranuni(0)*(&size-&number))+&number+1;
       set computer_generated point = k;
       if _error_ then abort;
       output;
     %end;
     stop;
   run;
```

```
    data boota;
      set boota (KEEP = x);
      censor = 1;
      proc sort data = boota;
      by x;

      proc append base=boot data=boota;
    run;

    /***Transform Data to Right-Censored***/

    data boot;
      set boot;
      x = 160 - x;
      proc sort data = boot;
      by x;
    run;

    /***Run Kaplan Meier***/

    PROC LIFETEST data = boot method = km;
      time x*censor(0);
      ods output Means=means&i;
    run;

   %end; /***END OF i = 100 Boot Samples***/

   %do j = 2 %to 100;
      proc append base = means1 data = means&j;
   %end;
   run;

%mend boottest; /***END OF Boottest Macro***/


/***Creating the UCLs***/

%macro test(size=, percent_Of_Detects=);

   /***100 UCLS Loop***/

   %DO t = 1 %to 100;

    data computer_generated;
      do z = 1 to &size;
      x = &beta*rangam(0,&alpha);
      output;
      end;
      proc sort data = computer_generated;
      by x;
    run;

    data computer_generated;
      set computer_generated (KEEP = x);
    run;

    /***Call the Boottest Macro***/
```

```
    %boottest(size=&sample, percent_Of_Detects=&detects)

    data means1;
      set means1;
      Mean = 160 - Mean;
      proc sort data = means1;
      by Mean;

    data means1 (RENAME = (Mean = UpperCL));
      set means1;

    /***95% UCL of 100 Boot Samples***/

    data UCL&t;
      set means1 (FIRSTOBS = 95 OBS=95);
    run;

    dm 'out;clear;log;clear;results;clear';

   %end; /***END OF 100 UCLs***/

   %do q = 2 %to 100;
     proc append base = UCL1 data = UCL&q;
   %end;

  run;

%mend test; /***End of Test Macro***/


%macro results;

   /***TEST DIFFERENT SAMPLE SIZE***/

   %DO sample = 10 %to 50 %BY 10;

      /***TEST DIFFERENT PERCENT OF NONDETECTS***/

      %DO detects = 10 %to 50 %BY 10;

        /***CALL TEST MACRO***/

        %test(size=&sample, percent_Of_Detects=&detects)

        data UCL1;
          set UCL1 (KEEP = UpperCL);
          PROC MEANS DATA = UCL1;
          var UpperCL;
          proc print data = UCL1;
          title "&sample" '_UCL_' "&detects";
        run;

        data gamma2.UCLs&sample&detects;
          set UCL1;
          count = 0;
```

```
          IF UpperCL > (&alpha*&beta) THEN count = 1;
          total + count;
          proc print data = gamma2.UCLs&sample&detects;
          title "&sample" '_UCL_' "&detects";

        data gamma2.Good_Bad_&sample&detects;
          set gamma2.UCLs&sample&detects (FIRSTOBS = 100 OBS = 100);
          IF total >= 95 THEN Result = 'Good';
          ELSE Result = 'BAD';
          proc print data = gamma2.Good_Bad_&sample&detects noobs;
          var Result total;
          title "&sample" '_UCL_Results' "&detects";
        run;

      %end; /***END OF NONDETECT LOOP***/

   %end; /***END OF SAMPLE SIZE LOOP***/

%mend results; /***END RESULTS MACRO***/

%results  /***CALL RESULTS MACRO***/

run;
```

BIBLIOGRAPHY

[1] Cleves, Mario A., Gould, William W., and Gutierrez, Roberto G. An Introduction to Survival Analysi Using Stat. College Station, Texas. Stata Corporation. 2004

[2] Farnham, Irene M., Singh, Ashok K., Stetzenback, Klaus J., and Johannesson, Kevin "Treatment of Nondetects in Multivariate Analysis of Groundwater Geochemistry Data." Chemometrics and Intelligent Laboratory Systems, Vol. 60, p. 265-281, Jan. 2002

[3] Helsel, Dennis R. Nondetects and Data Analysis: Statistics for Censored Environmental Data. New Jersey. John Wiley & Sons. 2005

[4] Kaplan, E. L., and Meier, P. "Nonparametric Estimation From Incomplete Observations." Journal of the American Statistical Association, Vol. 53, p. 457- 481, 1958

[5] Rubinstein, Reuven Y. Simulation and the Monte Carlo Method. New York. John Wiley & Sons. 1981

[6] SAS Institute Inc. 2004. Base SAS® 9.1 Procedures Guide. Cary, NC: SAS Institute Inc.

[7] Singh, Anita, Nocerino, John. "Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations." Chemometrics and Intelligent Laboratory Systems, Vol. 60, p. 69-86, Jan. 2002

[8] Smith, Peter J. Analysis of Failure and Survival Data. Boca Raton, Florida. Chapman & Hall/CRC. 2002

VITA

Graduate College
University of Nevada, Las Vegas

Violeta Graciela Hennessey

Local Address:
1329 Del Mar
Las Vegas, Nevada 89119

Home Address:
14702 Marklena Lane
Cypress, Texas 77429

Degree:
Bachelor of Science, Computer Science, 2003
Texas State University

Thesis Title: An Investigation of the Kaplan-Meier Upper Confidence Limit for the Population Mean From Environmental Samples with Nondetects

Thesis Examination Committee:
Chairperson, Dr. Ashok K. Singh, Ph.D.
Committee Member, Dr. Rohan Dalpatadu, Ph.D.
Committee Member, Dr. Hokwon Cho, Ph.D.
Graduate Faculty Representative, Dr. Laxmi Gewali, Ph.D.