A modified bootstrap method for handling non detects in environmental data

Michelle Melissa Guillaume

University of Nevada, Las Vegas

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

Repository Citation
https://digitalscholarship.unlv.edu/rtds/1830
A MODIFIED BOOTSTRAP METHOD FOR HANDLING
NON DETECTS IN ENVIRONMENTAL
DATA

by

Michelle Melissa Guillaume
Bachelor of Science
University of New Orleans, 2002

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
August 2005
INFORMATION TO USERS

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 bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send 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.
The Thesis prepared by
Michelle Guillaume

Entitled
A Modified Bootstrap Method for Handling Non-Detects in
Environmental Data

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
ABSTRACT

A Modified Bootstrap Method for Handling Non-Detects in Environmental Data

by

Michelle Melissa Guillaume

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

Many samples taken for environmental purposes have concentrations that are measured below the detection limit. Below the detection limit does not mean that there is no concentration present; there may be a concentration present that cannot be accurately measured by the analytical instrument or cannot be determined with 95% confidence that the concentration is greater than zero (EPA, 1989). These concentrations are defined non-detects in the data. When the concentration of the contaminant of concern in a sample falls below the detection limit, it is standard laboratory procedure to represent these data as <DL. The statistical procedure for handling non-detects in a data set depends on the probability distribution of the concentrations in the sample. There are several well accepted methods for handling non-detects in environmental data: Restricted Maximum Likelihood Estimation, Expectation-Maximization Method, and Cohen’s Adjustment to name a few (Singh, A., and Nocerino, J., 2001). All of these methods are parametric in nature, and require the knowledge of the form of the probability distribution function of the contaminant. In this thesis, a modified bootstrap method will be
developed and its performance evaluated for the treatment of non-detects in environmental data when the data is normally distributed. The proposed method can be easily adapted for non-normal populations.
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ iii

LIST OF TABLES ........................................................................................................................ vi

LIST OF FIGURES ........................................................................................................................ vii

CHAPTER 1  INTRODUCTION .................................................................................................. 1

CHAPTER 2 TESTING NORMALITY IN THE PRESENCE OF CENSORED DATA
.......................................................................................................................................................... 10
   Example 1A (Substitute DL/2) ................................................................................................... 11
   Example 1B (Probability Plot for Censored Data) ................................................................. 12
   Example 2A (Substitute Zero) ................................................................................................ 15
   Example 2B (Probability Plot for Censored Data) ................................................................. 17

CHAPTER 3 METHODOLOGY ............................................................................................... 20

CHAPTER 4 PROGRAMMING ............................................................................................... 25

CHAPTER 5 SIMULATIONS AND RESULTS ......................................................................... 29

CHAPTER 6 SUMMARY AND CONCLUSIONS .................................................................... 34

APPENDIX I .............................................................................................................................. 36

REFERENCES ............................................................................................................................. 42

VITA ............................................................................................................................................ 44
LIST OF FIGURES

FIGURE 1  Superfund Sites in the United States ................................................................. 3
FIGURE 2  Normal Probability Plot of $X_1^*$ From $N(50, 5)$ ........................................ 12
FIGURE 3  Normal Probability Plot for Censored Data $X_1^*$ ........................................ 15
FIGURE 4  Normal Probability Plot of $X_2^*$ From $N(50,5)$ ........................................... 16
FIGURE 5  Normal Probability Plot for Censored Data $X_2^*$ .......................................... 19
FIGURE 6  Coverage for Differing Sample Sizes (Modified Bootstrap Method) ............... 32
FIGURE 7  Coverage for Differing Sample Sizes (Winsorized Method) ............................ 32
FIGURE 8  Average 95% UCL for Differing Sample Sizes (Modified Bootstrap Method) ........ 33
FIGURE 9  Average 95% UCL for Differing Sample Sizes (Winsorized Method) .......... 33
ACKNOWLEDGEMENTS

I would like to take this opportunity to say thank you to some of the people that helped me with my accomplishments along this journey.

Thanks to:

❖ Dr. Singh who guided me along the way and always encouraged and supported me even when I doubted myself.

❖ My committee members: Dr. Laxmi P. Gewali, Dr. Rohan Dalpatadu, and Dr. Dieudonne D. Phanord.

❖ Norval and Linda without whom I would not be achieving this feat. Their love and support have helped me accomplished so much. For me, their guidance has been invaluable. I love you.
CHAPTER 1

INTRODUCTION

With all of the different issues concerning the United States today, hazardous waste and its cleanup may not be a top priority; however, the problem is much bigger than most people realize. There are certain sites around the United States that contain hazardous material of one form or another, but the Environmental Protection Agency (EPA) only qualifies the most highly contaminated sites as Superfund Sites. The Superfund program establishes a national program for responding to the releases of hazardous substances into the environment. There is a process involved in order to determine whether a site should be classified as a superfund site and placed on the National Priorities List (NPL). The following criteria in the next paragraph was referenced from (EPA, 1994).

First, a site is proposed to the NPL and the EPA then accepts public comments on the sites, responds to the comments, and places those sites that continue to meet the requirements for listing. Those requirements are three conditions of which the site needs to meet one: first is the EPA’s Hazardous Ranking System (HRS); this is the principal mechanism used by the EPA to place uncontrolled waste sites on the NPL. Scores for the sites are calculated using a structured analysis approach which assigns numerical values to factors that relate to risk based on conditions at the site. The second way a site may be placed on the NPL by the state in which the site lies. Each state in the U.S. is allowed to
place one site as a top-priority site regardless of the score that the site received from the HRS. In the event that the HRS score is low and another site has already been designated by the state as top-priority, there is a third way to get a site on the NPL; however, the site must meet all three of the following requirements: The Agency for Toxic Substances and Disease Registry (ATSDR) must issue a health advisory that recommends removing people from the site, the EPA must determine that the site poses a significant threat to public health and the EPA has to determine it will be more cost-effective to use its remedial authority than to use its emergency removal authority to respond to the site.

Currently, there are a little over one thousand three hundred Superfund sites scheduled for cleanup on the NPL in the United States (http://www.scorecard.org/env-releases/land/). In Nevada, there is only one site currently on the NPL, namely the Carson River Mercury Site. Its main contaminants include mercury, arsenic, and lead compounds. Even though there is only one NPL site in the state of Nevada, there are other areas of the country that have numerous sites on the NPL. For example, New Jersey has one hundred and sixteen sites on the NPL of which fifty have contaminated or threatened drinking water sources, and Pennsylvania has ninety five sites on the NPL of which fifty nine have contaminated or threatened drinking water sources. In fact, it is estimated that eleven million people in the United States, including 3-4 million children live within one mile of a federal Superfund site. The highest likelihood of exposure to any contaminants of concern is through contaminated water and ground soil. For a complete map of the Superfund Sites in the United States, please refer to Figure 1 on the next page (http://www.scorecard.org/env-releases/land/).
The EPA is continually collecting data from possibly infiltrated locations and trying to determine whether or not there is enough chemical concentration at the location to warrant a cleanup. This can be difficult because the instrument being used to detect traces of the chemical can only measure down to a certain amount. Anything below that amount, the instrument does not detect; however, that does not mean that a trace amount of the hazard is not present. Any measurements taken where this happens are considered below the detection limit of the instrument, are typed non-detects, and are labeled <DL.

DLs can be in a couple of forms. An Instrument Detection Limit (IDL) is generally the lowest amount of a substance that can be detected by an instrument; it is a measure only of the DL for the instrument, and does not consider any effects that sample matrix, handling, and preparation may have (EPA, 1989). A Method Detection Limit (MDL) is the minimum concentration of a substance that can be measured and reported with 95%
confidence that the value is greater than zero. The MDL can be calculated by analyzing a low-level real matrix sample, containing the contaminant at levels 2-10 times the expected detection limit. The formula is \( \text{MDL} = 2.896 \times \text{standard deviation of pooled nine results} \) (ME, 2004). In the previous equation, the sample standard deviation is being used. This is because the population standard deviation is never known.

Another measurement that can be calculated is the Practical Quantitation Limit (PQL). The PQL is the lowest level of quantitation that can be reliably achieved within specified limits of precision. For most environmental samples, the PQL is taken as five to ten times the MDL (ME, 2004). Two other methods for determining the PQL are available; they are analysis of Water Supply (WS) Performance Evaluation data (PE-data) and the Minimum Level (ML) MDL-Multiplier Approach. For information on how these methods work as well as the advantages and disadvantages of each method, please refer to the following: (EPA 1999, November 17).

The EPA is not required to report <DL measurements. A data set of measurements for a particular location may contain as few as 1% non-detects or as many as 90% non-detects. There are many ways that statisticians can choose to deal with this situation depending on the percentage of non-detects in the dataset in order to try and reliably predict whether a specific site should be cleaned or not. Independent of which method is being used, there needs to be an accurate computational procedure to calculate the mean concentration and related quantities necessary to evaluate attainment of the cleanup standard based on a random sample. The mean of the sampling data is an estimate of the mean contamination of the entire sample area, but does not convey information regarding the reliability of the estimate. It is possible, through the use of a confidence interval (CI)
to provide a range of values within which the true mean is located. An upper one-sided confidence interval should be used to test whether the site has attained the cleanup standard. The formula for calculating an upper one-sided CI for the mean, as given by the EPA is

\[
\mu_{\text{ua}} = \bar{x} + t_{1-\alpha, df} \cdot \frac{s}{\sqrt{n}}
\]

where \( \bar{x} \) is the computed mean level of contamination, and \( s \) is the corresponding standard deviation (EPA, February 1989a). The following rule should then be applied:

- if \( \mu_{\text{ua}} < Cs \), conclude area is clean
- if \( \mu_{\text{ua}} \geq Cs \), conclude area is dirty

\( Cs \) stands for the cleanup standards which are determined by EPA during site-specific endangerment assessments (EPA, February 1989a). The cost and timeliness of a cleanup are extensive.

We will now take a look at some data restricted by a DL along with some different procedures for how to handle such data.

Let’s consider the following data set with a detection limit of 75:

100, 94, 70, 64, 82, 86, 74, 98, 90, 80

If the EPA decides not to report any values below the DL, all values under 75 will be eliminated. The data set will become:

100, 94, 82, 86, 98, 90, 80

The data is now considered “truncated”, and in particular, type I left truncated. On the other hand, if the EPA does decide to report the values as below the DL, then the data becomes:

100, 94, 82, 86, 98, 90, 80
All of the original data values that were below 75 become undetected with our imposed DL; therefore, those values are represented in the above data set by <75. The data is now considered “censored”, and in particular, type I censored data. Censored data are classified into four major categories: truncated vs. censored, left vs. right, single vs. multiple and censored type I vs. censored type II (Cohen 1991, pp.3-5). In this thesis, we will be working with type I censored data as mentioned in the example above. Two main approaches involving several different methods may be employed to handle type I censored data.

The first approach is parametric which means that the form of the probability distribution function of the contaminant needs to be known. Methods that utilize this approach include Maximum Likelihood Estimation (MLE), Cohen’s Adjustment, and the Expectation-Maximization Algorithm (Cohen 1991, pp.3-5).

The MLE method is rigorous in nature. The environmental engineers usually use the Normal and Lognormal distributions, and then Maximum Likelihood equations are solved iteratively using some suitable numerical method such as the Newton-Raphson method. For larger data sets, the MLE method makes use of the number of data values above the DL, the proportion of data values below the DL, and the parametric form of the distribution. Maximum Likelihood Estimators are calculated by maximizing the likelihood function of the sample for the parameters mean, \( \mu \), and standard deviation, \( \sigma \). The MLE method is an efficient method for large enough data sets; however, Gleit (1985) found that the MLE method did not perform well for small sample sizes even when the assumed distribution was known.
The second parametric method comes from Cohen [1950, 1959], who derived the Maximum Likelihood (ML) equations for censored samples and prepared tables of the constants needed to obtain the MLE's of the mean and standard deviation with known distribution. Cohen's adjustment assumes that the detection limit is the same for each reported value for each individual constituent. A new mean and standard deviation are found using a series of steps utilizing Cohen's ML equations along with first-order bias correction terms tabulated by Saw [1961] and simplified by Schneider [1986]. Typically, Cohen's adjustment is used for data that contains less than 50% of the values being labeled as below the detection limit ((Gibbons, 1987) EPA, 1992c).

The final parametric method to be briefly introduced here for dealing with type I censored data is the EM Algorithm. The EM Algorithm is an algorithm that derives ML estimates of the parameters in a model based on incomplete data including missing values (Watanabe, 2004, p.7). Two major advantages to the EM Algorithm include not having to solve any major inverse matrix problem as well as the fact that computations following the calculation of the conditional expected value constitute estimations in cases where there are no missing values. i.e. the estimations can be performed easily without having to worry about a missingness pattern. (Watanabe, 2004, p.17). Gleit [1985] used this procedure for left-censored samples and found it to possess a lower Mean Squared Error (MSE) than the various other likelihood procedures. He recommended the use of the EM Algorithm which replaces all of the non-detects by the conditional expected value of the order statistics.

A second approach to dealing with censored data is non-parametric or distribution free. One of the most popular non-parametric methods used to handle censored data is
the substitution method. The most common substitutions made are to replace the values labeled <DL with DL, 0, or DL/2. Other than those three most common substitutions, some arbitrary fraction of DL is also used for some situations. Substituting 0 yields the smallest estimate of the mean along with the highest variation; on the other hand, substituting DL gives the highest estimate of the mean and the lowest estimation of the variance (Zhang, 2004). Farnham, Singh, Klaus, and Stetzenback (2002) found that substitution using DL/2 produces superior results compared to substitution with DL or 0; however, these latter substitutions are still employed in some cases. Some results for the above method include the following: Gilbert and Kinnison (1981) studied the methods of substitution and concluded that substituting for a detection limit is biased. A similar conclusion was drawn for the substitution for a detection limit by Gleit (1987). In general, performance of all substitution methods deteriorates when the number of non-detects in the data exceeds approximately 25% (Singh 2002). In an effort to make up for these findings, Gilbert (1987) considered ways to calculate an unbiased estimate of the sample mean. To accomplish this, there should be censored data from a normal or lognormal distribution. This data should be sorted with all of the values below the DL at the bottom of the distribution; these values are deleted along with the same amount of values at the top of the distribution being deleted. The value found by computing the mean of this new data is the “trimmed mean”.

The purpose of this thesis is to evaluate the performance of and compare two parametric methods for dealing with censored data when the data are normally distributed. They are a modified bootstrap method and a method known as “Winsorizing” the data. In the next chapter, normality plots for censored data will be
compared. Testing normality of the data is compulsory when dealing with censored data.

Both of the parametric methods mentioned above will be described in detail in chapter 3.

A description of the programming used to evaluate those two methods will then be given in chapter 4; there, it will be shown how our data sets and confidence intervals were calculated. Chapter 5 contains the tables and charts created from our simulations, and finally, a summary and conclusions will be given in the final chapter.
CHAPTER 2

TESTING NORMALITY IN THE PRESENCE OF CENSORED DATA

Approximate normality of the data distribution is typically assumed in computing a UCL of the mean. The traditional normal distribution based formulae for computing parametric UCL break down when the data distribution is moderately skewed to heavily skewed (Singh, et al., 2005). In this chapter, we use simulated data sets with non-detects and demonstrate the problems with running the standard test of normality in the presence of non-detects. We also illustrate a modified method for testing normality when some of the observations in the sample are <DL. The following are examples using simple substitution methods as well as a method for plotting data that contains the non-detects. All data was generated using MINITAB 14 software assuming a normal distribution with \( \mu = 50 \) and \( \sigma = 5 \).
Example 1A (Substitute DL/2)

A simulated data set of size 30 was generated from a normal distribution with μ = 50 and σ = 5, N (50, 5). The generated data values are displayed in Table 1.

<table>
<thead>
<tr>
<th>X1</th>
<th>41.9750</th>
<th>43.7731</th>
<th>45.5107</th>
<th>45.8328</th>
<th>45.9683</th>
<th>45.9817</th>
<th>46.7329</th>
<th>47.0154</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.7335</td>
<td>47.9031</td>
<td>47.9636</td>
<td>47.9807</td>
<td>48.5583</td>
<td>48.5982</td>
<td>50.5176</td>
<td>50.5334</td>
</tr>
<tr>
<td></td>
<td>50.9412</td>
<td>51.8424</td>
<td>51.8484</td>
<td>51.8552</td>
<td>52.1595</td>
<td>52.3907</td>
<td>54.0281</td>
<td>54.4386</td>
</tr>
<tr>
<td></td>
<td>54.7304</td>
<td>55.2542</td>
<td>55.9507</td>
<td>56.0280</td>
<td>56.4508</td>
<td>62.7642</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A DL of 45.6 was imposed on the data so that 10% of the data, or three values contained non-detects. A substitution of DL/2 or 22.8 was then made for those three values which had been labeled as non-detects. See Table 2.

<table>
<thead>
<tr>
<th>X1*</th>
<th>22.8</th>
<th>22.8</th>
<th>22.8</th>
<th>45.8328</th>
<th>45.9683</th>
<th>45.9817</th>
<th>46.7329</th>
<th>47.0154</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.7335</td>
<td>47.9031</td>
<td>47.9636</td>
<td>47.9807</td>
<td>48.5583</td>
<td>48.5982</td>
<td>50.5176</td>
<td>50.5334</td>
</tr>
<tr>
<td></td>
<td>50.9412</td>
<td>51.8424</td>
<td>51.8484</td>
<td>51.8552</td>
<td>52.1595</td>
<td>52.3907</td>
<td>54.0281</td>
<td>54.4386</td>
</tr>
<tr>
<td></td>
<td>54.7304</td>
<td>55.2542</td>
<td>55.9507</td>
<td>56.0280</td>
<td>56.4508</td>
<td>62.7642</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A normal probability plot of the data was generated and displayed in figure 2. By looking at the graph, after the substitution, this data fails the standard normality test,
(P < 0.01), even though the data set was generated from a normal population.

Figure 2: Normal Probability Plot of X1* from N(50,5)

Example 1B (Probability Plot for Censored Data)

We will now use a modified normality test on the same data generated in Example 1A; however, instead of using a simple substitution for the values that are considered below the detection limit, the non-detects are kept in the data set as shown in table 3 on the following page.
Table 3: Generated Data N (50, 5) Containing Non-Detects

<table>
<thead>
<tr>
<th>X1</th>
<th>&lt; DL</th>
<th>&lt; DL</th>
<th>&lt; DL</th>
<th>45.8328</th>
<th>45.9683</th>
<th>45.9817</th>
<th>46.7329</th>
<th>47.0154</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.7335</td>
<td>47.9031</td>
<td>47.9636</td>
<td>47.9807</td>
<td>48.5583</td>
<td>48.5982</td>
<td>50.5176</td>
<td>50.5334</td>
</tr>
<tr>
<td></td>
<td>50.9412</td>
<td>51.8424</td>
<td>51.8484</td>
<td>51.8552</td>
<td>52.1595</td>
<td>52.3907</td>
<td>54.0281</td>
<td>54.4386</td>
</tr>
<tr>
<td></td>
<td>54.7304</td>
<td>55.2542</td>
<td>55.9507</td>
<td>56.0280</td>
<td>56.4508</td>
<td>62.7642</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard tests of normality cannot be used in the presence of non-detects. Instead, we used the method of probability plotting for censored data (Zacks, 1992). An explanation of this method is given in the following steps:

1. If \( m \) = the number of non-detect values in the data set of \( n \) values, then we will plot the points \((x_i, z_i)\) where \( i = m+1, m+2, \ldots, n \). Referring to Table 5 above, the \( x_i \)'s are the uncensored data values sorted in ascending order. The \( z_i \)'s are the corresponding normal scores found using the following equation:

\[
    z_i = \Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right)
\]  

where \( \Phi^{-1}(\cdot) \) is the inverse cumulative distribution function (cdf) of the standard normal distribution.

2. Using the calculated pairs of points, fit a least-squares regression line.

\[
    x_i = a + bz_i
\]
Then, the estimated mean of the normal distribution is given by the intercept, $a$, and the standard deviation is estimated by the slope, $b$.

The data along with the calculation results are displayed in Table 4 below.

Table 4: Computational Results and Points for Fitted Line Plot Using X1

<table>
<thead>
<tr>
<th>Index</th>
<th>$i - \frac{3/8}{n + 1/4}$</th>
<th>$z_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.119835</td>
<td>-1.17581</td>
<td>45.8328</td>
</tr>
<tr>
<td>5</td>
<td>.152893</td>
<td>-1.02411</td>
<td>45.9683</td>
</tr>
<tr>
<td>6</td>
<td>.185950</td>
<td>-0.89292</td>
<td>45.9817</td>
</tr>
<tr>
<td>7</td>
<td>.219008</td>
<td>-0.77555</td>
<td>46.7329</td>
</tr>
<tr>
<td>8</td>
<td>.252066</td>
<td>-0.66800</td>
<td>47.0154</td>
</tr>
<tr>
<td>9</td>
<td>.285124</td>
<td>-0.56769</td>
<td>47.7335</td>
</tr>
<tr>
<td>10</td>
<td>.318182</td>
<td>-0.47279</td>
<td>47.9031</td>
</tr>
<tr>
<td>11</td>
<td>.351240</td>
<td>-0.38198</td>
<td>47.9636</td>
</tr>
<tr>
<td>12</td>
<td>.384298</td>
<td>-0.29421</td>
<td>47.9807</td>
</tr>
<tr>
<td>13</td>
<td>.417355</td>
<td>-0.20866</td>
<td>48.5583</td>
</tr>
<tr>
<td>14</td>
<td>.450413</td>
<td>-0.12462</td>
<td>48.5982</td>
</tr>
<tr>
<td>15</td>
<td>.483471</td>
<td>-0.04144</td>
<td>50.5176</td>
</tr>
<tr>
<td>16</td>
<td>.516529</td>
<td>0.04144</td>
<td>50.5334</td>
</tr>
<tr>
<td>17</td>
<td>.549587</td>
<td>0.12462</td>
<td>50.9412</td>
</tr>
<tr>
<td>18</td>
<td>.582645</td>
<td>0.20866</td>
<td>51.8424</td>
</tr>
<tr>
<td>19</td>
<td>.615702</td>
<td>0.29421</td>
<td>51.8484</td>
</tr>
<tr>
<td>20</td>
<td>.648760</td>
<td>0.38198</td>
<td>51.8552</td>
</tr>
<tr>
<td>21</td>
<td>.681818</td>
<td>0.47279</td>
<td>52.1595</td>
</tr>
<tr>
<td>22</td>
<td>.714876</td>
<td>0.56769</td>
<td>52.3907</td>
</tr>
<tr>
<td>23</td>
<td>.747934</td>
<td>0.66800</td>
<td>54.0281</td>
</tr>
<tr>
<td>24</td>
<td>.780992</td>
<td>0.77555</td>
<td>54.4386</td>
</tr>
<tr>
<td>25</td>
<td>.814050</td>
<td>0.89292</td>
<td>54.7304</td>
</tr>
<tr>
<td>26</td>
<td>.847107</td>
<td>1.02411</td>
<td>55.2542</td>
</tr>
<tr>
<td>27</td>
<td>.880165</td>
<td>1.17581</td>
<td>55.9507</td>
</tr>
<tr>
<td>28</td>
<td>.913223</td>
<td>1.36087</td>
<td>56.028</td>
</tr>
<tr>
<td>29</td>
<td>.946281</td>
<td>1.60982</td>
<td>56.4508</td>
</tr>
<tr>
<td>30</td>
<td>.979339</td>
<td>2.04028</td>
<td>62.7642</td>
</tr>
</tbody>
</table>
A least squares regression line based on equation (9) was fitted to the data from columns three and four in Table 4, and the fitted line plot using those pairs of points is displayed in Figure 3.

\[ x_i = 50.29 + 4.849 z_i \]

![Fitted Line Plot](image)

Figure 3: Normality Probability Plot for Censored Data X1*

In Figure 3, the data appears normal with an estimated mean 50.29 and standard deviation 4.849.

**Example 2A (Substitute Zero)**

The following example follows a similar substitution method as Example 1A. The data was generated from the same distribution, N(50, 5), and a DL of 44 was imposed to ensure 10% of the data, or three values, were non-detect values. This time, a substitution of 0 was made for those three values, and a normal probability plot was generated. Refer to Tables 5-6 and Figure 4 on the next page.
Table 5: Generated Data X2 from N(50, 5)

<table>
<thead>
<tr>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.4633</td>
</tr>
<tr>
<td>42.6097</td>
</tr>
<tr>
<td>43.4625</td>
</tr>
<tr>
<td>44.2967</td>
</tr>
<tr>
<td>44.4413</td>
</tr>
<tr>
<td>44.9118</td>
</tr>
<tr>
<td>45.6203</td>
</tr>
<tr>
<td>45.7301</td>
</tr>
<tr>
<td>45.8569</td>
</tr>
<tr>
<td>46.4779</td>
</tr>
<tr>
<td>46.6243</td>
</tr>
<tr>
<td>46.6684</td>
</tr>
<tr>
<td>46.7602</td>
</tr>
<tr>
<td>48.2100</td>
</tr>
<tr>
<td>48.9528</td>
</tr>
<tr>
<td>49.0537</td>
</tr>
<tr>
<td>49.4611</td>
</tr>
<tr>
<td>49.7776</td>
</tr>
<tr>
<td>50.5183</td>
</tr>
<tr>
<td>50.6417</td>
</tr>
<tr>
<td>51.7323</td>
</tr>
<tr>
<td>51.7345</td>
</tr>
<tr>
<td>52.7263</td>
</tr>
<tr>
<td>53.8882</td>
</tr>
<tr>
<td>55.4460</td>
</tr>
<tr>
<td>55.8174</td>
</tr>
<tr>
<td>55.8702</td>
</tr>
<tr>
<td>57.2004</td>
</tr>
<tr>
<td>58.6596</td>
</tr>
<tr>
<td>59.5284</td>
</tr>
</tbody>
</table>

Table 6: Generated Data X2* with Substitution of 0

<table>
<thead>
<tr>
<th>X2*</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>44.2967</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44.4413</td>
<td>44.9118</td>
<td>45.6203</td>
<td>45.7301</td>
</tr>
<tr>
<td></td>
<td>45.8569</td>
<td>46.4779</td>
<td>46.6243</td>
<td>46.6684</td>
</tr>
<tr>
<td></td>
<td>46.7602</td>
<td>48.2100</td>
<td>48.9528</td>
<td>49.0537</td>
</tr>
<tr>
<td></td>
<td>49.4611</td>
<td>49.7776</td>
<td>50.5183</td>
<td>50.6417</td>
</tr>
<tr>
<td></td>
<td>51.7323</td>
<td>51.7345</td>
<td>52.7263</td>
<td>53.8882</td>
</tr>
<tr>
<td></td>
<td>55.4460</td>
<td>55.8174</td>
<td>55.8702</td>
<td>57.2004</td>
</tr>
<tr>
<td></td>
<td>58.6596</td>
<td>59.5284</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probability Plot of DL=44

- Normal
- Mean: 45.22
- StdDev: 15.92
- N: 30
- KS: 0.377
- P-Value <0.010

Figure 4: Normal Probability Plot of X2* From N(50,5)
Once again, with a substitution of 0, the data fails the standard normality test, \( P < .01 \), even though the data was generated from a normal population.

**Example 2B (Probability Plot for Censored Data)**

Now we will illustrate the modified normality test again for the same data used in Example 2A. The same procedures will be used as were described in Example 1B. First, the data containing the non-detects is displayed in Table 7.

**Table 7: Generated Data N(50, 5) Containing Non-Detects**

<table>
<thead>
<tr>
<th>X2</th>
<th>&lt; DL</th>
<th>&lt; DL</th>
<th>&lt; DL</th>
<th>44.2967</th>
<th>44.4413</th>
<th>44.9118</th>
<th>45.6203</th>
<th>45.7301</th>
</tr>
</thead>
<tbody>
<tr>
<td>45.8569</td>
<td>46.4779</td>
<td>46.6243</td>
<td>46.6684</td>
<td>46.7602</td>
<td>48.2100</td>
<td>48.9528</td>
<td>49.0537</td>
<td></td>
</tr>
<tr>
<td>49.4611</td>
<td>49.7776</td>
<td>50.5183</td>
<td>50.6417</td>
<td>51.7323</td>
<td>51.7345</td>
<td>52.7263</td>
<td>53.8882</td>
<td></td>
</tr>
<tr>
<td>55.4460</td>
<td>55.8174</td>
<td>55.8702</td>
<td>57.2004</td>
<td>58.6596</td>
<td>59.5284</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Following the modified method, we imposed equation (8) on the data, and created Table 8 on the following page.

**Table 8: Computational Results and Points for Fitted Line Plot Using X2**

17
Table 8: Computational Results and Points for Fitted Line Plot Using X2

<table>
<thead>
<tr>
<th>Index</th>
<th>$\left( i - \frac{3}{8} \right) / \left( \frac{n + 1/4}{x_i} \right)$</th>
<th>$z_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.119835</td>
<td>-1.17581</td>
<td>44.2967</td>
</tr>
<tr>
<td>5</td>
<td>.152893</td>
<td>-1.02411</td>
<td>44.4413</td>
</tr>
<tr>
<td>6</td>
<td>.185950</td>
<td>-0.89292</td>
<td>44.9118</td>
</tr>
<tr>
<td>7</td>
<td>.219008</td>
<td>-0.77555</td>
<td>45.6203</td>
</tr>
<tr>
<td>8</td>
<td>.252066</td>
<td>-0.66800</td>
<td>45.7301</td>
</tr>
<tr>
<td>9</td>
<td>.285124</td>
<td>-0.56769</td>
<td>45.8569</td>
</tr>
<tr>
<td>10</td>
<td>.318182</td>
<td>-0.47279</td>
<td>46.4779</td>
</tr>
<tr>
<td>11</td>
<td>.351240</td>
<td>-0.38198</td>
<td>46.6243</td>
</tr>
<tr>
<td>12</td>
<td>.384298</td>
<td>-0.29421</td>
<td>46.6684</td>
</tr>
<tr>
<td>13</td>
<td>.417355</td>
<td>-0.20866</td>
<td>46.7602</td>
</tr>
<tr>
<td>14</td>
<td>.450413</td>
<td>-0.12462</td>
<td>48.21</td>
</tr>
<tr>
<td>15</td>
<td>.483471</td>
<td>-0.04144</td>
<td>48.9528</td>
</tr>
<tr>
<td>16</td>
<td>.516529</td>
<td>0.04144</td>
<td>49.0537</td>
</tr>
<tr>
<td>17</td>
<td>.549587</td>
<td>0.12462</td>
<td>49.4611</td>
</tr>
<tr>
<td>18</td>
<td>.582645</td>
<td>0.20866</td>
<td>49.7776</td>
</tr>
<tr>
<td>19</td>
<td>.615702</td>
<td>0.29421</td>
<td>50.5183</td>
</tr>
<tr>
<td>20</td>
<td>.648760</td>
<td>0.38198</td>
<td>50.6417</td>
</tr>
<tr>
<td>21</td>
<td>.681818</td>
<td>0.47279</td>
<td>51.7323</td>
</tr>
<tr>
<td>22</td>
<td>.714876</td>
<td>0.56769</td>
<td>51.7345</td>
</tr>
<tr>
<td>23</td>
<td>.747934</td>
<td>0.66800</td>
<td>52.7263</td>
</tr>
<tr>
<td>24</td>
<td>.780992</td>
<td>0.77555</td>
<td>53.8882</td>
</tr>
<tr>
<td>25</td>
<td>.814050</td>
<td>0.89292</td>
<td>55.446</td>
</tr>
<tr>
<td>26</td>
<td>.847107</td>
<td>1.02411</td>
<td>55.8174</td>
</tr>
<tr>
<td>27</td>
<td>.880165</td>
<td>1.17581</td>
<td>55.8702</td>
</tr>
<tr>
<td>28</td>
<td>.913223</td>
<td>1.36087</td>
<td>57.2044</td>
</tr>
<tr>
<td>29</td>
<td>.946281</td>
<td>1.60982</td>
<td>58.6596</td>
</tr>
<tr>
<td>30</td>
<td>.979339</td>
<td>2.04028</td>
<td>59.5284</td>
</tr>
</tbody>
</table>

The least squares regression line based on equation (9) was fitted to the data from columns three and four in Table 8, and the fitted line plot using those pairs of points is displayed in Figure 5.
Once again, the data does appear normal with an estimated mean 49.24 and standard deviation 5.416.

The above two examples show that:

(1) The standard test of normality gives incorrect results (rejects normality when a sample with non-detects is generated from a normal distribution).

(2) The modified test of normality based on the Q-Q plot for censored samples works well.
CHAPTER 3

METHODOLOGY

In this section, details will be given to describe the two methods that will be evaluated and compared in this thesis. These methods are a modified bootstrap method and a method known as "winsorizing" the data.

The bootstrap method is a fairly new resampling technique. It requires the use of modern computer power in order to make inferences about a data set. This thesis utilizes a modified non-parametric bootstrap technique in order to develop estimates for 95% confidence intervals for the mean. The key idea is to resample from the original data to create replicate datasets, from which the variability of the quantities of interest can be assessed without long-winded and error-prone analytical calculation [Davison, 1997].

The bootstrap method has both a parametric and a non-parametric approach. Even though we will be focusing on the parametric bootstrap approach for our work in this thesis, we will make a brief description of the non-parametric bootstrap for the sake of completeness.

For any bootstrap resampling technique, one must calculate the standard error of the bootstrap estimates. The standard error for the non-parametric bootstrap is found differently from the parametric bootstrap in that it is not based on information derived from a parametric model of the data. The parametric bootstrap assumes the population to
come from a certain distribution, say normal, exponential, or lognormal whereas the non-parametric bootstrap takes a distribution-free approach. After the standard error is calculated, the process for using the non-parametric bootstrap is the same that it would be for the parametric bootstrap which will be discussed in detail shortly. The parametric bootstrap can be used to find more accurate answers to textbook formulas, and can provide answers in problems for which no textbook formulae exist [Efron, 1993].

As mentioned above, this thesis uses a modified parametric bootstrap technique. This technique is reliant on analyzing bootstrap samples; so let us first look at what a bootstrap sample contains. Say we have a sample $x = (x_1, x_2, ..., x_n)$ containing $n$ data values. A single bootstrap sample, $x^{*1}$, will consist of values drawn randomly from our original sample, $x$, with replacement until we have $n$ bootstrap data points $(x^*_1, x^*_2, ..., x^*_n)$. Because we are sampling with replacement from our original data set, some points from the original data may appear once, twice, or not at all in our bootstrap sample. For resampling, it is recommended to have a large number of bootstrap samples; in this thesis, we will draw 1000 bootstrap samples. Once we have 1000 bootstrap samples from our original data, we can then calculate the estimate of the unknown parameter for each bootstrap sample and then compute the standard error of the estimate. An algorithm for estimating the standard error of the estimate is given by Efron [1993] and is given below.

First, we select $B$ independent bootstrap samples, $x^{*1}, x^{*2}, ..., x^{*B}$, each consisting of $n$ data values drawn with replacement from $x$. In this case, $B = 1000$. Second, we evaluate the bootstrap estimate corresponding to each bootstrap sample. For example, if we wanted to estimate the standard error of the mean, we would first find the mean of each bootstrap sample $\bar{x}_1, \bar{x}_2, ..., \bar{x}_B \quad b = 1, 2, ..., B$. Finally, the standard error of the mean
can be estimated by the sample standard deviation of the B bootstrap replications. In the case of the standard error of the mean again, we would calculate it based on the following formula:

\[
est(SE^\mu) = \sqrt{\frac{\sum_{b=1}^{B} [\bar{x}(b) - \bar{x}(\cdot)]^2}{B - 1}} \quad \text{where} \quad \bar{x}(\cdot) = \frac{\sum_{b=1}^{B} \bar{x}(b)}{B}
\]  

(2)

Once the standard error has been calculated, bootstrap confidence intervals can be obtained. A bootstrap-t interval for the mean can be found using the following method. For each bootstrap sample, compute

\[
U^*(b) = \frac{\bar{x}(b) - \bar{x}}{\est(se^*(b))}
\]

(3)

where \( \bar{x}(b) \) is the value of the mean for the bootstrap sample, \( x^b \), \( \bar{x} \) is the mean of the original data, \( x \), and \( \est(se^*(b)) \) is the estimated standard error for the bootstrap sample.

The \( \alpha th \) percentile of \( U^*(b) \) is estimated by \( y(\alpha) \) such that

\[
\frac{\{U^*(b) \leq y(\alpha)\}}{B} = \alpha
\]

(4)

Finally, the "bootstrap-t" confidence interval for \( \bar{x} \) is given by
The above approach for the parametric bootstrap standard error and confidence intervals is a typical one. The approach used here is a modified version of the above methodology, using a percentile bootstrap method and will be explained thoroughly in the programming chapter.

The second method that we will be evaluating the performance of is known as “winsorizing” the data. This technique was investigated by Dixon and Tukey (1968). The mean of a data set can be heavily influenced by extreme values in the tails of the distribution. Winsorizing the data compensates for this by setting the tail values equal to a certain percentile value. For our purposes in evaluating data containing differing percentages of non-detects, using this method, the data is sorted and then values on both ends are replaced according to the percentage of non-detects in the data. For example, let us take the following sorted data set to be the one we need to analyze:

\(< 4, < 4, 5, 5, 7, 8, 9, 10, 13, 20\)

The above data set has \(n = 10\) data values of which two are labeled as below the detection limit. So 20% of the data has non-detect values. In this case, we replace 20% of the values on both ends of the data. The two non-detect values will be replaced by the next largest datum and the two largest values in the set will be replaced by the next smallest datum. In general, the first \(k\) smallest values are replaced by the \((k+1)\)th smallest value, \(x_{(k+1)}\), and the first \(k\) largest values are replaced by the \((k+1)\)th largest value, \(x_{(n-k)}\). Applying this to the above data set would give us the new “winsorized” data set:

\(5, 5, 5, 5, 7, 8, 9, 10, 10, 10\)
We can now compute the sample mean, $\bar{x}$, and standard deviation, $s$. The Winsorized mean, $\bar{x}_w$, is an unbiased estimator of $\mu$ (Gilbert, 1987). The Winsorized standard deviation is

$$s_w = \frac{s(n-1)}{v-1}$$

which is an approximately unbiased estimator for $\sigma$ where $n$ is the total number of data values, and $v$ is the number of data not replaced during the Winsorization. When the data are from a normal distribution, which will be our assumption here, the upper and lower limits of a two-sided 100(1-$\alpha$)% confidence interval about the mean are

$$\bar{x}_w \pm t_{1-\alpha/2, n-1} \frac{s_w}{\sqrt{n}}$$

(Gilbert, 1987).

For this thesis, we evaluate the performance of this method using data sets from a normal distribution with differing numbers of data values as well as different percentages of non-detects contained within the data.
CHAPTER 4

PROGRAMMING

All of the programming for this thesis was done using software called R. R is available as Free Software and is similar to the S programming language. R is currently being widely used and it provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. In order to download the R software, go to the following website: http://www.r-project.org/. On the left side, under download, click on CRAN (Comprehensive R Archive Network). Scroll down the main page and pick any one of the links for the USA. In the precompiled binary distributions box, find the link that is compatible with your operating system. From here, you will choose the base package and then click on rw2011.exe-this is the file that you will download. Once it is saved and installed to your computer, R is ready to use. If there are any problems, you can refer to or download the manual under the documentation heading on the left side of either your download page or the r-project homepage. It was highly effective for this project in dealing with and analyzing censored data.

The program code used for this project is included at the back of the paper in APPENDIX I. The code works in the following manner:
(1) It initially generates \( n \) values from a standard normal distribution with mean \( \mu \) and standard deviation \( \sigma \). A detection limit is imposed depending on the number and percentage of values <DL that we want to appear in the data.

(2) A Winsorized data set is created by replacing the number of values <DL as well as replacing that same number of values from the upper end of the data the same way it was explained in chapter 2. The mean, \( \hat{\mu}_w \), and standard deviation, \( s_w \), of the Winsorized data are computed along with the standard error.

(3) Next, a new estimate of the standard deviation is computed using the fact that there are \( m \) observations <DL. Using this fact, the ratio of non-detects to the number of data values is the following:

\[
\frac{m}{n} = P(X < DL) = \Phi\left(\frac{DL - \mu}{\sigma}\right)
\]  

where \( \Phi(\cdot) = \text{cdf of N}(0,1) \) distribution. Solving for \( \sigma \) gives us the following formula for the new estimate of the standard deviation:

\[
\hat{\sigma}_1 = \frac{DL - \hat{\mu}_w}{qnorm\left[\Phi^{-1}\left(\frac{m}{n}\right)\right]}
\]

(4) In this step, we generate \( m \) random numbers from the normal distribution \( N(\hat{\mu}_w, \hat{\sigma}_1) \) truncated above by the detection limit (DL). This is achieved by using the probability integral transform:
\[ X \sim F(x) \Rightarrow F(x) \sim U(0,1) \]  

(12)

Where,

\[ F(x) = \Phi\left( \frac{x - \hat{\mu}_w}{\hat{\sigma}_1} \right) \Phi\left( \frac{DL - \hat{\mu}_w}{\hat{\sigma}_1} \right) \]  

(13)

To generate an observation from F(x), we first generate \( u \sim U(0,1) \), and then compute

\[ x = F^{-1}(u) \]. Using (13) solve for \( x \):

\[ \Phi\left( \frac{x - \hat{\mu}_w}{\hat{\sigma}_1} \right) = F(x) \cdot \Phi\left( \frac{DL - \hat{\mu}_w}{\hat{\sigma}_1} \right) \]

\[ \frac{x - \hat{\mu}_w}{\hat{\sigma}_1} = \Phi^{-1} \left[ u \cdot \Phi\left( \frac{DL - \hat{\mu}_w}{\hat{\sigma}_1} \right) \right] \]

\[ x = \hat{\mu}_w + \hat{\sigma}_1 \cdot \Phi^{-1} \left[ u \cdot \Phi\left( \frac{DL - \hat{\mu}_w}{\hat{\sigma}_1} \right) \right] \]  

(14)

This leads to the following equation:

\[ xc[i] = \mu_w + \hat{\sigma}_1 \cdot \text{qnorm}\left(u[i]\right) \cdot \text{pnorm}\left(\frac{DL - \hat{\mu}_w}{\hat{\sigma}_1}\right) \quad \text{where } i = 1,2,\ldots,m \]  

(15)
(5) When we have the completed data set, the mean is calculated. We now have a Winsorized data set as well as a competed data set which we have found using a modified method.

(6) Bootstrap resampling is then applied, the upper confidence limits (UCL) are calculated for the Winsorized and modified data, and a counter is established to count how many times the UCL-nu and UCL_W are greater than the true mean of 100. Finally, the average UCL_boot and average UCL_W are calculated.

The printed results show n (the sample size), m (number of values <DL), true mean (100), true sd (10), niter (1000 bootstrap samples are taken), ucl_counter_nu (number of times our modified data is greater than the true mean), ucl_counter_w (number of times Winsorized data is greater than the true mean), ucl_nu (95% UCL for modified data), ucl_w (95% UCL for Winsorized data), avg. ucl_nu (average UCL after 1000 replications for modified data), avg. ucl_w (average UCL after 1000 replications for Winsorized data). We used $\mu = 100$ and $\sigma = 10$ in this thesis.
CHAPTER 5

SIMULATIONS

Simulations were done using five different sample sizes: \( n = 20, n = 40, n = 60, n = 80, \) and \( n = 100, \) and \( d\% \) non-detects in the sample: \( d = 5, \ldots, 45 \) at increments of 5%.

For each combination of sample size with a certain percentage of non detects one thousand iterations were run in order to estimate the percentage of times the mean was greater than the true mean, 100, a 95% Upper Confidence Limit (UCL), and an average UCL for each of the two methods under examination. Each of the above computations was done for both methods under examination: the modified bootstrap method as well as the Winsorized method. After all percentages of non detects were run for a certain sample size, a table was constructed to display the results. Tables 1-5 show these results.

Column one shows the specific percentage of non detects used for that simulation followed by column two which displays exactly how many non detects were created for that sample size in order to achieve the correct percentage. Columns three and four contain the number of times counted where each of the means for the two methods was higher that the true mean of one hundred. This number is out of one thousand iterations. The modified bootstrap counter is represented as Ucl_counter_nu and the Winsorized counter is represented by Ucl_counter_W. Next comes columns five and six which show 95% UCL for each method with Ucl_nu for the modified bootstrap method and Ucl_W
for the Winsorized method. Finally, Columns seven and eight contain the average UCL for each method with Avg. Ucl_boot for the modified bootstrap method and Avg. Ucl_W for the Winsorized method.

Table 9: Results Obtained for n = 20 (Modified Bootstrap and Winsorized Methods)

<table>
<thead>
<tr>
<th>%ND</th>
<th>m</th>
<th>Ucl_counter_nu</th>
<th>Ucl_counter_W</th>
<th>Ucl_nu</th>
<th>Ucl_W</th>
<th>Avg. Ucl_boot</th>
<th>Avg. Ucl_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>1</td>
<td>966</td>
<td>952</td>
<td></td>
<td>103.5304</td>
<td>103.7843</td>
<td>104.2332</td>
</tr>
<tr>
<td>10%</td>
<td>2</td>
<td>962</td>
<td>950</td>
<td></td>
<td>103.8144</td>
<td>103.7006</td>
<td>104.2319</td>
</tr>
<tr>
<td>15%</td>
<td>3</td>
<td>964</td>
<td>958</td>
<td></td>
<td>103.9186</td>
<td>103.9</td>
<td>104.2766</td>
</tr>
<tr>
<td>20%</td>
<td>4</td>
<td>956</td>
<td>959</td>
<td></td>
<td>104.4113</td>
<td>103.6723</td>
<td>104.2757</td>
</tr>
<tr>
<td>25%</td>
<td>5</td>
<td>958</td>
<td>959</td>
<td></td>
<td>104.3911</td>
<td>103.7137</td>
<td>104.1764</td>
</tr>
<tr>
<td>30%</td>
<td>6</td>
<td>949</td>
<td>953</td>
<td></td>
<td>105.4149</td>
<td>103.3008</td>
<td>104.2890</td>
</tr>
<tr>
<td>35%</td>
<td>7</td>
<td>902</td>
<td>935</td>
<td></td>
<td>105.2638</td>
<td>103.9595</td>
<td>103.8128</td>
</tr>
<tr>
<td>40%</td>
<td>8</td>
<td>899</td>
<td>951</td>
<td></td>
<td>109.1654</td>
<td>104.4609</td>
<td>104.6913</td>
</tr>
<tr>
<td>45%</td>
<td>9</td>
<td>652</td>
<td>943</td>
<td></td>
<td>107.8269</td>
<td>104.1530</td>
<td>101.9230</td>
</tr>
</tbody>
</table>

Table 10: Results Obtained for n = 40 (Modified Bootstrap and Winsorized Methods)

<table>
<thead>
<tr>
<th>%ND</th>
<th>m</th>
<th>Ucl_counter_nu</th>
<th>Ucl_counter_W</th>
<th>Ucl_nu</th>
<th>Ucl_W</th>
<th>Avg. Ucl_boot</th>
<th>Avg. Ucl_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>2</td>
<td>976</td>
<td>955</td>
<td></td>
<td>102.7294</td>
<td>102.7913</td>
<td>103.2002</td>
</tr>
<tr>
<td>10%</td>
<td>4</td>
<td>971</td>
<td>961</td>
<td></td>
<td>102.6154</td>
<td>102.6556</td>
<td>102.9854</td>
</tr>
<tr>
<td>15%</td>
<td>6</td>
<td>976</td>
<td>960</td>
<td></td>
<td>102.9993</td>
<td>102.9271</td>
<td>103.1490</td>
</tr>
<tr>
<td>20%</td>
<td>8</td>
<td>972</td>
<td>950</td>
<td></td>
<td>103.0959</td>
<td>102.8234</td>
<td>103.0980</td>
</tr>
<tr>
<td>25%</td>
<td>10</td>
<td>961</td>
<td>957</td>
<td></td>
<td>103.4173</td>
<td>102.9351</td>
<td>103.0839</td>
</tr>
<tr>
<td>30%</td>
<td>12</td>
<td>978</td>
<td>961</td>
<td></td>
<td>105.2495</td>
<td>102.9664</td>
<td>103.9659</td>
</tr>
<tr>
<td>35%</td>
<td>14</td>
<td>836</td>
<td>954</td>
<td></td>
<td>103.2188</td>
<td>103.1143</td>
<td>102.3303</td>
</tr>
<tr>
<td>40%</td>
<td>16</td>
<td>772</td>
<td>942</td>
<td></td>
<td>104.3566</td>
<td>103.1146</td>
<td>102.1454</td>
</tr>
<tr>
<td>45%</td>
<td>18</td>
<td>842</td>
<td>936</td>
<td></td>
<td>113.2324</td>
<td>102.9044</td>
<td>105.5940</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
### Table 11: Results Obtained for n = 60 (Modified Bootstrap and Winsorized Methods)

<table>
<thead>
<tr>
<th>%ND</th>
<th>m</th>
<th>Ucl_counter_num</th>
<th>Ucl_counter_W</th>
<th>Ucl_num</th>
<th>Ucl_W</th>
<th>Avg. Ucl_boot</th>
<th>Avg. Ucl_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>3</td>
<td>974</td>
<td>957</td>
<td>102.2900</td>
<td>102.3130</td>
<td>102.5741</td>
<td>102.2858</td>
</tr>
<tr>
<td>10%</td>
<td>6</td>
<td>965</td>
<td>948</td>
<td>102.1488</td>
<td>102.1284</td>
<td>102.4912</td>
<td>102.2435</td>
</tr>
<tr>
<td>15%</td>
<td>9</td>
<td>964</td>
<td>951</td>
<td>102.2271</td>
<td>102.2002</td>
<td>102.4978</td>
<td>102.2840</td>
</tr>
<tr>
<td>20%</td>
<td>12</td>
<td>955</td>
<td>946</td>
<td>102.4138</td>
<td>102.3737</td>
<td>102.3716</td>
<td>102.2908</td>
</tr>
<tr>
<td>25%</td>
<td>15</td>
<td>948</td>
<td>956</td>
<td>102.8204</td>
<td>102.4817</td>
<td>102.4659</td>
<td>102.5216</td>
</tr>
<tr>
<td>30%</td>
<td>18</td>
<td>912</td>
<td>961</td>
<td>102.4867</td>
<td>102.3835</td>
<td>102.1948</td>
<td>102.5231</td>
</tr>
<tr>
<td>35%</td>
<td>21</td>
<td>869</td>
<td>941</td>
<td>103.2996</td>
<td>102.4034</td>
<td>102.1511</td>
<td>102.5669</td>
</tr>
<tr>
<td>40%</td>
<td>24</td>
<td>678</td>
<td>947</td>
<td>102.6977</td>
<td>102.3691</td>
<td>101.1374</td>
<td>102.7814</td>
</tr>
<tr>
<td>45%</td>
<td>27</td>
<td>831</td>
<td>954</td>
<td>111.3729</td>
<td>102.5940</td>
<td>104.3783</td>
<td>103.2103</td>
</tr>
</tbody>
</table>

### Table 12: Results Obtained for n = 80 (Modified Bootstrap and Winsorized Methods)

<table>
<thead>
<tr>
<th>%ND</th>
<th>m</th>
<th>Ucl_counter_num</th>
<th>Ucl_counter_W</th>
<th>Ucl_num</th>
<th>Ucl_W</th>
<th>Avg. Ucl_boot</th>
<th>Avg. Ucl_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>4</td>
<td>970</td>
<td>955</td>
<td>101.7407</td>
<td>101.7726</td>
<td>102.2126</td>
<td>101.9413</td>
</tr>
<tr>
<td>10%</td>
<td>8</td>
<td>968</td>
<td>949</td>
<td>101.9517</td>
<td>101.858</td>
<td>102.1633</td>
<td>101.9261</td>
</tr>
<tr>
<td>15%</td>
<td>12</td>
<td>975</td>
<td>965</td>
<td>101.9145</td>
<td>101.8509</td>
<td>102.1988</td>
<td>101.9824</td>
</tr>
<tr>
<td>20%</td>
<td>16</td>
<td>967</td>
<td>958</td>
<td>102.3552</td>
<td>101.9623</td>
<td>102.2584</td>
<td>102.0541</td>
</tr>
<tr>
<td>25%</td>
<td>20</td>
<td>952</td>
<td>961</td>
<td>102.2349</td>
<td>101.9016</td>
<td>102.1351</td>
<td>102.0325</td>
</tr>
<tr>
<td>30%</td>
<td>24</td>
<td>938</td>
<td>950</td>
<td>102.6391</td>
<td>101.9736</td>
<td>102.2125</td>
<td>102.1328</td>
</tr>
<tr>
<td>35%</td>
<td>28</td>
<td>919</td>
<td>952</td>
<td>103.2337</td>
<td>101.9833</td>
<td>102.2069</td>
<td>102.1393</td>
</tr>
<tr>
<td>40%</td>
<td>32</td>
<td>939</td>
<td>952</td>
<td>105.5666</td>
<td>101.9516</td>
<td>103.2041</td>
<td>102.3023</td>
</tr>
<tr>
<td>45%</td>
<td>36</td>
<td>946</td>
<td>950</td>
<td>113.6347</td>
<td>102.1234</td>
<td>107.5152</td>
<td>102.5590</td>
</tr>
</tbody>
</table>

### Table 13: Results Obtained for n = 100 (Modified Bootstrap and Winsorized Methods)

<table>
<thead>
<tr>
<th>%ND</th>
<th>m</th>
<th>Ucl_counter_num</th>
<th>Ucl_counter_W</th>
<th>Ucl_num</th>
<th>Ucl_W</th>
<th>Avg. Ucl_boot</th>
<th>Avg. Ucl_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>5</td>
<td>970</td>
<td>948</td>
<td>101.6999</td>
<td>101.6778</td>
<td>101.9628</td>
<td>101.7116</td>
</tr>
<tr>
<td>10%</td>
<td>10</td>
<td>974</td>
<td>960</td>
<td>101.6903</td>
<td>101.6681</td>
<td>101.9492</td>
<td>101.7327</td>
</tr>
<tr>
<td>15%</td>
<td>15</td>
<td>964</td>
<td>951</td>
<td>101.7188</td>
<td>101.5688</td>
<td>101.9046</td>
<td>101.6852</td>
</tr>
<tr>
<td>20%</td>
<td>20</td>
<td>967</td>
<td>941</td>
<td>102.1505</td>
<td>101.7616</td>
<td>102.0573</td>
<td>101.7670</td>
</tr>
<tr>
<td>25%</td>
<td>25</td>
<td>926</td>
<td>954</td>
<td>101.8165</td>
<td>101.7685</td>
<td>101.7754</td>
<td>101.8503</td>
</tr>
<tr>
<td>30%</td>
<td>30</td>
<td>932</td>
<td>960</td>
<td>102.5032</td>
<td>101.8015</td>
<td>101.9793</td>
<td>101.9284</td>
</tr>
<tr>
<td>35%</td>
<td>35</td>
<td>904</td>
<td>962</td>
<td>102.9467</td>
<td>101.9320</td>
<td>101.9294</td>
<td>102.0046</td>
</tr>
<tr>
<td>40%</td>
<td>40</td>
<td>868</td>
<td>955</td>
<td>103.7800</td>
<td>101.8624</td>
<td>102.0923</td>
<td>102.0379</td>
</tr>
<tr>
<td>45%</td>
<td>45</td>
<td>396</td>
<td>946</td>
<td>103.0511</td>
<td>101.9698</td>
<td>98.9981</td>
<td>102.3074</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The results displayed in Tables 9-13 are graphed in Figures 6-9.

Figure 6: Coverage for Differing Sample Sizes (Modified Bootstrap Method)

Figure 7: Coverage for Differing Sample Sizes (Winsorized Method)
Figure 8: Average 95% UCL for Differing Sample Sizes (Modified Bootstrap Method)

Figure 9: Average 95% UCL for Differing Sample Sizes (Winsorized Method)
CHAPTER 6

SUMMARY AND CONCLUSIONS

The characterization of superfund sites in the United States is an extremely important problem for the EPA’s Superfund Program as it places contaminated sites on the NPL. Cleanup of a site can be extremely costly, and it is very important that the EPA correctly identify which sites need to be cleaned using the strictest of criteria. However, when data from the site comes back with a significant number of data values labeled as <DL, the choice of statistical methods for data analysis is not clear. Several methods to deal with this were discussed in this thesis; however only two of them were evaluated on the basis of performance. Those were a modified bootstrap method and the Winsorized method.

In this thesis, a program was created using the R programming language in order to evaluate the performance of a newly created Modified Bootstrap Method as well as a previously created method known as Winsorization as they relate to handling non-detects in environmental data. Data from a normal distribution was generated for all the simulations done in this thesis. The simulations were performed and analyzed, and we were able to create the following charts and come up with some conclusions. Tables 1-5 and Figures 5 and 6 showed that the Modified Bootstrap Method has good coverage for all sample sizes up until around 30% of the data contain non-detect values. The most extreme drops in coverage take place when the sample sizes are very large,
n = 100, and very small, n = 20. The Winsorized method, on the other hand, performs well for up to 45% non-detects in the sample and has excellent coverage for all sample sizes. This is an important result for the Winsorized Method.

Figures 7 and 8 show that the average UCL's are comparable for the two methods.

Future projects which could come out of this thesis include the following:

(1) Showing how the modified bootstrap created in this thesis can easily be adapted to other parametric distributions.

(2) Investigating how the Winsorized mean performs for non-normal populations.

(3) Investigation of the performance of the Q-Q-plot method, shown in chapter 3 of this thesis, for estimating the mean and the standard deviation.
APPENDIX I

R PROGRAMMING CODE USED INCLUDING COMMENTS DENOTED BY A PRECEDING # SYMBOL

# truemean = true mean
# truesd = true sd
# initialize the true mean to be 100

truemean <- 100
# initialize the true standard deviation to be 10

truesd <- 10
# set the number of non-detects to be present in the data (changes for each simulation)

m<-4
# creating a variable m1 to be used in calculation of the detection limit

m1<-m+1
# choosing the number of data values (for this thesis 20,40,60,80,100)

n<-40
# setting the number of iterations to perform in a variable called niter

niter<-1000
# creating a variable to use when the program winsorizes the data (takes from the top)

nm1<-n-m+1
# initializing several arrays to hold data and perform calculations

u<-array(c(1,20))
a<-array(c(1,40))
b<-array(c(1,40))
x<-array(c(1,40))
xx<-array(c(1,40))
xw<-array(c(1,40))
xc<-array(c(1,40))

# outputting to a file where all of the results will be stored
sink("outal.txt", append = T)

# setting the counters for both the modified bootstrap and winsor methods to zero
ucl_counter_nu<-0

ucl_counter_w<-0

# labeling so that the resulting data is clear
print('n, m, true mean, true sd, niter')

# the following will be used to check in the results to see that we are using the right initial values
print(n)

print(m)

print(truemean)

print(truesd)

print(niter)

# printing a separator within the results to make them more clear
print("="*49)

# BEGIN ITERATION LOOP
for(iter in 1:niter)
{
    # print('iteration #')
    # print(iter)
    # creating a dataset, x, randomly from a normal distribution with 40 data values
    x<-rnorm(40, truemean, truesd)
    # print('complete data')
    # print(x)
    # sorting the data in ascending order
    x<-sort(x)
    # print('sorted complete data')
    # print(x)
    # calculating a detection limit for the data created
    DL<-(x[m]+x[m1])/2
    # print('DL')
    # print(DL)
    # xw = Winsorized data
    # putting our created data into the xw array created to hold the winsorized data
    for (i in 1:40){xw[i] <- x[i]}
    # the next two steps replace values at the bottom and the top of the dataset
    for (j in 1:m) {xw[j] <- x[m1]}
    for (k in n-n1:n) {xw[k] <- x[n-m]}
    # print('winsorized data')
    # print(xw)
    # wmn, wsd = Winorized mean and sd
    # computing the mean of the newly created winsorized dataset
    wmn[iter] <- mean(xw)
    # computing the standard deviation of the newly created winsorized dataset
    wsd<-sd(xw)
# wsd2 = Winsorized sum of squared deviations

wsd2 <- (n-l)*sd(xw)*sd(xw)

# computing the standard error for the winsorized mean

σδ_çµν<-(ν-1)*σ;ρτ(σδ2)/((ν-2*μ-1)*σ;ρτ(ν*(ν-1)))

# print('winsorized mean, winsorized sd, stderr(wins. mean')
# print(wmn[iter])
# print(wsd)
# print(sd_wmn)
# print('winsorized sd, winsorized stderr')

# nusd = estimated sd using the fact that there are m observations < DL

nusd<-(DL-wmn)/qnorm(m/n)

# Next, complete the data set x

# generating m values from a uniform distribution with mean 0 and sd 1

u<-runif(m,0,1)

# replacing the non-detect values with the values found after using the following equation

for (i in l:m){xc[i]<-wmn+nusd*qnorm(u[i]*pnorm((DL-wmn)/nusd))}

for (i in m+1:40){xc[i]<-x[i]}

# print('completed data xc')
# print(xc)

# mnc = mean(completed x)

mnc[iter]<-mean(xc)

# initializing variable B to 1000

B<-1000

# starting the bootstrap resampling with replacement

for (j in 1:B)
{
    xx<-sample(xc,replace=T)
    # mean of the completed dataset

    mnstar[j]<-mean(xx)
}

# standard deviation of the mean of the completed dataset

bootsd<-sd(mnstar)

# print('sd(mnstar) from bootstrap')
# print(bootsd)

# calculating upper confidence limit for our modified bootstrap
ucl_nu[iter]<-mnc[iter]+1.96*bootsd
# calculating a value for t utilizing the number of data values and the number of non
detects
t<-qt(.95, df=n-2*m-l)

# print('t for winsorized mean')
# print(t)
# calculating upper confidence limit for winsorized data
ucl_w[iter]<-wmn[iter] + t*sd_wmn

# print('ucl_nu, ucl Winsorized')
# print(ucl_nu[iter])
# print(ucl_w[iter])
# setting the conditions under which the counters should increase for both methods
if (ucl_nu[iter] > truemean) ucl_counter_nu <- ucl_counter_nu + 1
if (ucl_w[iter] > truemean) ucl_counter_w <- ucl_counter_w + 1

} # END ITER LOOP
# calculating the mean for each of the ucls for each method
mean_uclb<-mean(ucl_nu)
mean_uclw<-mean(ucl_w)

# print('95% ucls using bootstrap and winsorized methods')
# sorting both sets of data
mnc<-sort(mnc)
wmn<-sort(wmn)

# print(mnc)
# print(wmn)
# creating labels for clear output as well as displaying the values below each label
print('ucl_counter_nu')
print(ucl_counter_nu)
print('ucl_counter_w')
print(ucl_counter_w)
print(mnc[950])
print(wmn[950])
print('average bootstrap ucl, winsorized ucl')
print(mean_uclb)
print(mean_uclw)
# just printing a separator in the results to distinguish between runs of the program
print('==================================')

**NOTE**: The print commands preceded by a # symbol are not necessary. They are for double checking purposes and can clog up the program if used for all one thousand iterations.
REFERENCES


EPA (Environmental Protection Agency) (1994). National Oil and Hazardous Substances Pollution Contingency Plan. Part 300, Subpart E.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Newman, M.C., Greene, K.D., & Dixon, P.M. 1995. Uncensor v.4.0. Savannah River Ecology Laboratory. 91


VITA

Graduate College
University of Nevada, Las Vegas

Michelle Melissa Guillaume

Local Address:
1895 N. Green Valley Pkwy. #1921
Henderson, NV 89074

Degrees:
Bachelor of Science, Mathematics, 2002
University of New Orleans

Thesis Title: A Modified Bootstrap Method for Handling Non Detects in Environmental Data.

Thesis Examination Committee:
Chairperson, Dr. A.K. Singh Ph.D.
Committee Member, Dr. Dieudonne Phanord, Ph.D.
Committee Member, Dr. Rohan Dalpatadu, Ph.D.
Graduate Faculty Representative, Dr. Gewali, Ph.D.