1-1-2005

Bayesian estimation of the normal mean in the presence of non-detects

Ahmed Khago
University of Nevada, Las Vegas

BAYESIAN ESTIMATION OF THE NORMAL MEAN IN THE PRESENCE OF NON-DETECTS

by

Ahmed Khago

Bachelor of Science
University of Wyoming
1999

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
May 2005

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

March 11, 2005

The Thesis prepared by

Ahmed Khago

Entitled

Bayesian Estimation of The Normal Mean In Presence of non-Detects

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

Master of Science in Mathematical Sciences

Ashon Kumar Sen
Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

Bayesian Estimation of the Normal Mean In the Presence of Non-Detects

by

Ahmed Khago

Dr. Ashok Singh, Examination Committee Chair
Professor of Mathematics
University of Nevada, Las Vegas

This paper is concerned with the Bayesian approach to estimate the mean when encountered with left-censored data sets. Considering the joint non-informative prior, we derived the posterior probability density function of the mean of left-censored data. However, this density function is not recognizable and we can not analytically integrate it to obtain the normalizing constant. In other words, we can not compute analytically the posterior pdf or posterior moments. Numerical integration involving the adaptive Simpson quadrature rule was used in Mat-lab to obtain the posterior mean and the upper credible limit (UCL). Several numerical examples are given which illustrate the practical application of these results.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I'd like to thank my advisor, Dr. Singh for his dedication and care for his students and for giving me the opportunity to work and learn from him. I couldn't do it without you. I also would like to thank Dr. Dennis Murphy, Dr. Dan Asera and Rina Santos for their help and support and the committee for their positive feed back. Finally, I would like to thank my wife Nancy for her support, encouragement, and patience throughout this journey.

CHAPTER 1

INTRODUCTION

In environmental applications, many data measurements such as herbicide concentrations in soil, air, and water do not get reported because such measurements fall below a certain detection limit ("DL") and many groundwater monitoring applications of the United States Environmental Protection Agency ("EPA") do not require reporting such data. These measurements, however, cannot be ignored since they impact the upper confidence limit ("UCL") of the mean which is required for many remediation decisions. The DL of an analytical method is the lowest level of concentration of any particular substance that can be reliably detected and is statistically different from a "blank" reading.

In most environmental applications, this non-reported data, with observations recorded as being below a certain limit, is called "censored" data – which means the observations are not available at one or both ends. Censoring usually occurs when the pollutant concentration is very near or below the DL; however, this practice creates special problems and makes it difficult to analyze and summarize data sets and could lead

to biased estimations of the population parameters, such as the mean and the standard deviation.

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). Environmental Science applications mostly deal with type I left censored.

A data sample is said to be left truncated if the truncation point ("T") is known and the value of the observations below T is deleted or not reported, but the values above T are known and are reported. For example, consider the data set: 3, 4, 3, 5, 4, 3, <2, 2, 3, <2, <2, with a DL of 2. All the data values reported as "<2" will be eliminated and if no indication of how many observation were excluded, this would be called a type I, left—truncated sample. On the other hand, a data set of size "n" is said to be left—censored if the censoring level T is known and the value of the observations below T level is known (k observations) only to fall below T while the known observations above T level are fully known and reported (n-k observations). For the above example, the values reported as "<2" will not be eliminated.

The difference between truncated data and censored data is that the censored data points are those whose measured values are not known precisely, but are known to fall below some DL. On the other hand, truncated data points are those which are missing from the sample altogether due to sensitivity limits.

The most common method of dealing with censored data in environmental applications is the substitution method. One way is to delete the censored data. The reason behind the use of this method is that interest is always in the detected data. This method produces biased results because the statistics are computed for the detected data. A second way is to replace each censored observation with an arbitrary fraction of the DL. The most common substitution is to replace the censored data by zero, half the detection limit or by the detection limit itself. Singh and Nocerino (2002) pointed out the replacement of half the detection limit produced biased estimate of mean and error increases when multiple detection limits are present.

Another approach is the maximum likelihood estimation ("MLE"), which is often used in environmental studies. There are three types of information needed to perform the calculation: the values of data above detection limits, the proportion of data below detection limits, and the parametric form of the assumed distribution. For small data sets, however, the MLE would perform poorly (Gleit, 1985; Shumway, et al, 2002). MLE is an efficient method to estimate the parameters when the number of observations is large enough. MLE is obtained by maximizing the likelihood function ("L") for the parameters μ and σ.

The third approach involves non-parametric procedures, which are called the distribution-free methods and are commonly used in

environmental sciences. These methods are useful for censored data because they use the available information.

Researchers from various disciplines have studied the estimation of the parameters of the normal populations from censored samples. One of these researchers is Cohen [1950-1959]. He derived the maximum likelihood estimate ("MLE") of the mean and the standard deviation from censored data. Cohen's MLE uses both the detected observations and the proportion of data set below detection limits to compute statistics for the entire data set. The MLE method requires that the distribution of the data be known and specified. In environmental sciences, the normal and lognormal distributions are usually used. The MLE equation is then solved using numerical methods, such as the Newton-Raphson method; however, the MLE method has been shown to perform poorly with data set containing 25 to 50 observations (Gleit, 1985; Shumway, et al, 2002).

Gilbert and Kinnison (1981) studied and evaluated the methods of substitution, deleting censored data and Cohen's table lookup. They concluded that substituting for a detection limit is biased. Gleit (1985) found MLE did not perform well for a small data set, even though the assumed distribution is known. He concluded MLE methods work poorly for small sample sizes and the substitution method of detection limits also worked poorly. Gillion and Hesel (1986) found that the MLE method worked well when the assumed distribution matched that of data. They

also found that the substitution method worked poorly. Gilbert (1987) considered several methods to calculate an unbiased estimate of the sample mean. The data set should be from normal or lognormal distributions and should include censored data. The data set then should be sorted out and with an equal number of observations from both ends be deleted. The trimmed mean can then be calculated from these values. The trimmed mean is usually recommended to estimate the mean of a symmetric distribution, even if the data set does not have missing values.

Another method is called "winsorizing" the data set and is considered by Dixon and Tukey (1968), in which we replace the sorted data set at both ends of the data series with the next extreme value at both ends and compute the mean of the new data. The difference between the trimmed mean method and the winsorized method is the trimmed method discards data on both ends of the data set and computes the mean of the remaining data; but the winsorized method replaces data in both ends with the next most extreme datum in each end and then computes the mean of the new data set. Winsorization can be used to estimate the mean and the standard deviation of a symmetric distribution, even though the data set has missing values at one or both ends of the ordered data set.

Our goal in this paper is to compute the Bayes estimate of the mean of a normal population when the data set has non-detects. We

present several examples using simulated data and compute the Bayes estimate obtained from left-censored samples with that obtained from the uncensored samples.

The Bayesian method is a statistical method in which the parameters of the particular distribution are estimated based on the posterior distribution. Unlike the classical method, in the Bayesian method, the parameters are viewed as random variables. We start with what is called the "prior distribution"' which reflects the experimenter's prior believe about the population parameter θ. The statistician observes the sample from $f(x|\theta)$, the conditional pdf of the random variable X, given the random parameter θ. The Bayes theorem is then used to compute the posterior pdf of θ, given the sample. The posterior pdf is used to compute the Bayes estimate of θ, or a UCL for θ.

CHAPTER 2

METHODS

In this section, we give details of the popular methods to estimate the mean and variance of a population when only censored data is available. Some of these methods are the trimmed mean, the winsorized mean, and the maximum likelihood.

In situations where non-detect values are reported even when the measurements are below the detection limit, the population mean μ and the variance σ^2 can be estimated by calculating the sample mean $\bar{x}$ and the sample variance s^2 using one of the following :

1. Calculate $\bar{x}$ and s^2 using the full data set, including non-detect values.
2. Delete all non-detects and calculate only $\bar{x}$ and s^2 using only the detected data set.
3. Replace every non-detect values with zero and then calculate $\bar{x}$ and s^2.
4. Replace the non detected values with values generated from uniform over $[0, DL]$ then calculate $\bar{x}$ and s^2.

All of the methods mentioned above are known to be biased.

The Trimmed Mean

The trimmed mean is one of the methods of estimating the mean of a symmetric distribution and it's a compromise between the median and the mean. Several of the lowest and the highest observations are trimmed off (np observations), where $0 < p < .5$, and then the mean of what is left off, $(n(1-2p))$,is calculated. Common trimming is 25% of the data at each end. The resulting mean of the central 50% of data is commonly called the "trimmed mean." For example, suppose n=25, data collected from a symmetric distribution has a true mean μ. We can estimate μ using a 25% trimmed mean. We first compute $.25n$= .25(25) = 6.25. Hence, we can discard the 6 smallest and the 6 largest data. The mean of the remaining is 25-12=13; data is the estimate of the mean.

The Winsorized Mean

The use of the winsorized mean method is also one of the recommended methods to estimate the mean of censored data of symmetric distribution. Details of this method are given by Dixon and Tukey (1968).

Given a sample of size n, with k non-detect values, the winsorized procedure is described below:

1. Replace the k non-detects values by the next datum.

2. Replace the k largest values by the next smallest datum.
3. Calculate the sample mean $\bar{x}_w$ and standard deviation s_w of resulting n data.
4. The resulting estimate $\bar{x}_w$ is known to be an unbiased estimator of μ.

The following sample from a well represents the concentration for hazardous chemicals ordered from the smallest to the largest. Trace, trace, trace, .67, 2.4, 3.1, 3.5, 3.9, 4.1, 4.6, 5.7, 6.9, 7.5, and 9.1. Replace the three trace concentrations by .67 and the three largest concentrations by 5.7. The data becomes .67, .67, .67, .67, 3.1, 3.5, 3.9, 4.1, 4.6, 5.7, 5.7, 5.7, and 5.7. The sample mean of the new data is $\bar{x}_w$=3.36. This $\bar{x}_w$ is the winsorized mean.

Bootstrap Method

Bootstrap methods are non-parametric methods that require no assumptions regarding the population distribution such as the normality assumption. These methods are used to reduce the bias in point estimate and build a confidence interval for any parameter. It's a form of a larger class of methods that resample from the original data set and therefore are called resampling procedures. We can obtain accurate confidence intervals without having to make normal theory assumption and estimate the distribution directly from the data set. The procedure is described below:

Let $x_1, x_2, ..., x_n$ be a random sample of size n, then B bootstrap samples are generated from the original data set. Each bootstrap sample should have n elements, which is generated by sampling with replacement n times. Bootstrap replicates $\bar{X}_1, \bar{X}_2, ..., \bar{X}_b$ are calculated from the bootstrap samples. We next calculate the bootstrap standard error, $s_B = \sqrt{\frac{\sum(\bar{X}_i - \bar{X}_B)^2}{n-1}}$, and obtain the confidence interval using s_B. Finally, $(1-\alpha)100\%$ confidence interval for θ is $\left(\hat{\theta} - z_\alpha s_B, \hat{\theta} + z_\alpha s_B\right)$.

CHAPTER 3

One of the most difficult problems in environmental data analysis is deciding on the appropriate method of incorporating the censored data in computing summary statistics, corresponding tests of hypotheses, and interval estimation of parameters. This is mostly because the choice of method depends on the degree of censoring (for example, 10% versus 90% non-detects) and this also depends on the form of the probability distribution. Most of the commonly used methods involve replacing the detection limit with an arbitrary constant. This paper will provide a Bayesian estimate of the mean from left censored data set.

Left-truncated normal distribution has been utilized by a variety of disciplines, such as environmental sciences, economics and finance. Pearson and Lee (1908), Fisher (3), Hald (1949), and Cohen studied singly truncated normal samples when the truncation point is known and the sample size of unmeasured observations is unknown. Stevens (1938), Cochran (1949) and Hald (1949) studied singly truncated normal samples when the truncation point is known and the sample size of unmeasured observation is known.

Consider a random variable X from a normal distribution with a probability density function $f(x)$ specified as:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty \quad (1)$$

The Posterior Density of the Mean of Censored Data

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal distribution $N(\mu, \sigma)$ and suppose k of these measurements falls below the detection limit, DL. Let ϕ be the probability density function ("pdf") and Φ be the cumulative density function ("cdf"), then the likelihood function is the following (Persson and Rootzen 1977):

$$L(\underline{x}, \mu, \sigma) = [\Phi(z)]^k (2\pi\sigma)^{-(n-k)/2} \exp -[\sum_{i=k+1}^{n} (y_i + z\sigma)^2 / 2\sigma^2] \quad (2)$$

$\Phi(z)$= cumulative distribution function of the standard normal distribution, N(0,1), and k= the number of observations below detection limit.

$$\Phi(x) = p(X \leq DL) = \int_{-\infty}^{DL} f(x)\, dx \quad (3)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{DL} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2} dx \tag{4}$$

$\Phi(z)$ can be written as the cumulative density function:

$$[\Phi(z)] = \left\{ P\left(\frac{X-\mu}{\sigma} \leq \frac{DL-\mu}{\sigma} \right) \right\} \tag{5}$$

Let $Z = \frac{X-\mu}{\sigma}$, $dz = \frac{1}{\sigma} dx$, the following formula is obtained:

$$[\Phi(z)] = \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-\frac{z^2}{2}} dz \right\}, \quad \xi = \frac{DL-\mu}{\sigma} \tag{6}$$

The following graph shows the left-censored normal distribution :

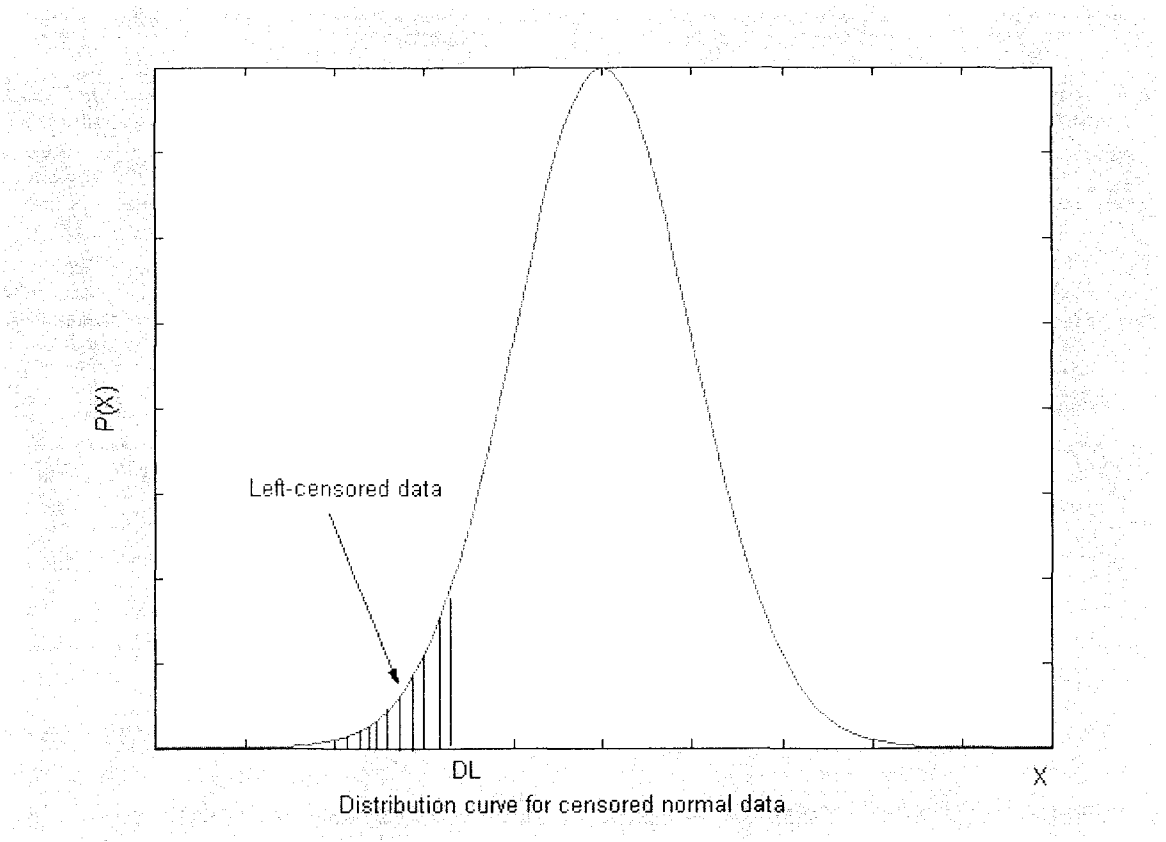

Figure 1: Censored normal distribution

$$[\Phi(z)]^k = \left\{1 - \frac{1}{\sqrt{2\pi}} \int_{\xi}^{\infty} e^{-\frac{z^2}{2}} dz\right\}^k \tag{7}$$

The integration of the normal distribution is easier using what is called the error function. The error function is twice the integral of the standardized normal distribution with $\mu = 0$ and $\sigma = 1$. The error function is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-u^2} du \tag{8}$$

The complementary error function is given by:

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-u^2} du \tag{9}$$

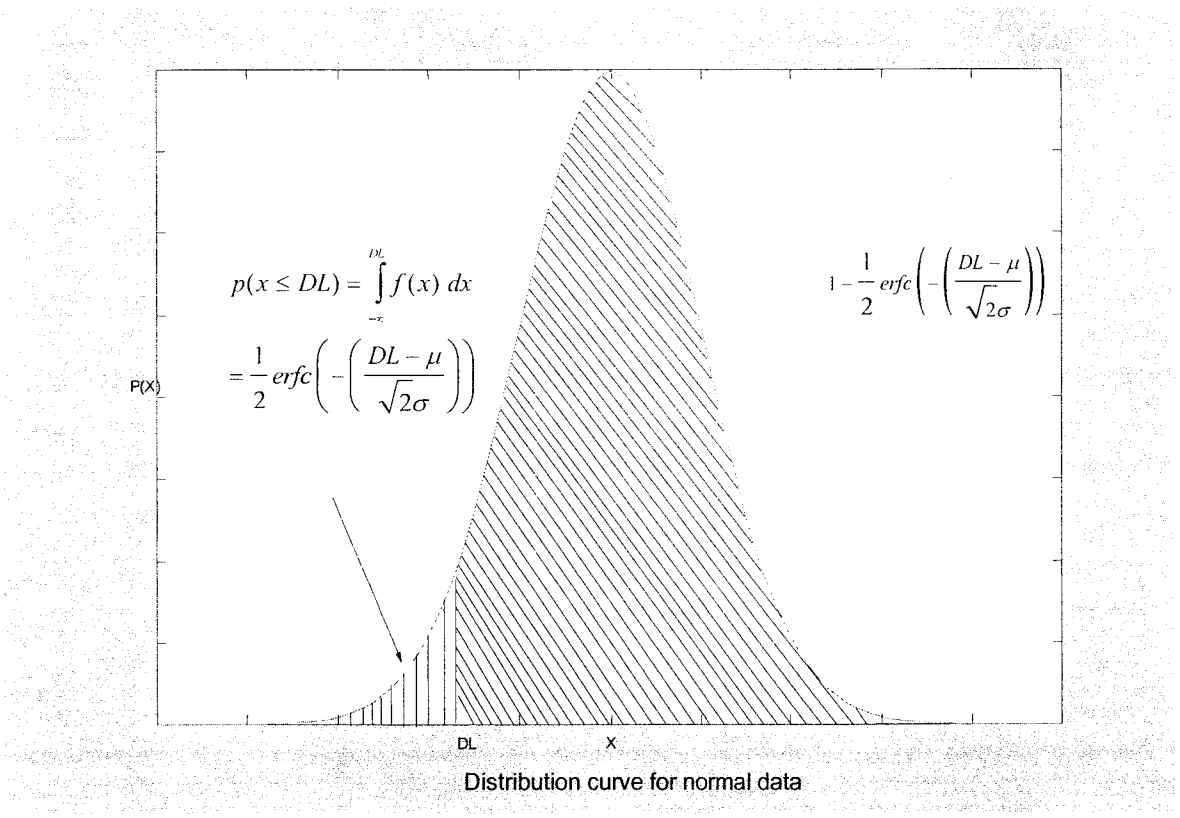

Figure: 2 Normal distribution in term of the error Function

As prior for μ and σ, we will use Jeffrey's the non-informative prior, (Jeffreys,1961):

$$g(\mu,\sigma) = 1/\sigma^2 \tag{10}$$

The posterior distribution is obtained from (2) and (10) via the Bayes theorem (Lee, 1989).

$$g^*(\mu,\sigma \mid \underline{x}) = \frac{f(\underline{x} \mid \mu,\sigma)g(\mu,\sigma)}{\iint f(\underline{x} \mid \mu,\sigma)g(\mu,\sigma)d\mu d\sigma} = \frac{f(\underline{x} \mid \mu,\sigma)g(\mu,\sigma)}{K(x)} \tag{11}$$

$K(x)$ =The Marginal distribution which is constant and free of μ and σ.

Which after substitution of the various terms, becomes:

$$g^*(\mu,\sigma \mid x) = \frac{[\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp-[\sum_{i=k+1}^{n} (x_i-\mu)^2/2\sigma^2]}{\iint [\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp-[\sum_{i=k+1}^{n} (x_i-\mu)^2/2\sigma^2]\, d\mu\, d\sigma} \tag{12}$$

The posterior could be written as:

$$g^*(\mu,\sigma \mid \underline{x}) \propto f(\underline{x} \mid \mu,\sigma)g(\mu,\sigma) \tag{13}$$

$$g^*(\mu,\sigma \mid x) \propto [\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp-[\sum_{i=k+1}^{n} (x_i-\mu)^2/2\sigma^2] \tag{14}$$

$$g^*(\mu \mid x) \propto \int_0^{\infty} [\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp-[\sum_{i=k+1}^{n} (x_i-\mu)^2/2\sigma^2]\, d\sigma \tag{15}$$

It is not possible to analytically integrate out μ and σ from this posterior density function, $g^*(\mu,\sigma \mid \underline{x})$, to obtain the marginal posterior pdf's: $g^*(\mu \mid x)$ and $g^*(\sigma^2 \mid x)$. Also, the conditional posterior pdf's: $g^*(\mu \mid x)$ and $g^*(\sigma^2 \mid x)$ are not recognizable densities. In other words, we do not know analytically the constant $K(x)$, such that $g^*(\mu \mid x)/K(x)$ is a properly normalized density, i.e. such that $\int g^*(\mu \mid x)dx = 1$.

The posterior distribution for μ

$$g^*(\mu \mid x) \propto [\Phi(z)]^k \exp -[n(\mu - \bar{x})^2 / 2\sigma^2] \tag{16}$$

The posterior truncated mean is given by

$$\hat{\mu}_B = E(x) = \int_{-\infty}^{\infty} x\, g^*(\mu \mid x)\, dx \tag{17}$$

$$g^*(\mu \mid x) \propto \int_0^{\infty} [\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp -[\sum_{i=k+1}^{n} (x_i - \mu)^2 / 2\sigma^2]\, dx \tag{18}$$

The posterior density of the mean of uncensored data

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal distribution $N(\mu,\sigma)$, then the likelihood function is the following:

$$L(x,\mu,\sigma) = (2\pi\sigma)^{-n/2} \exp-[\sum_{i=1}^{n}(y_i + z\sigma)^2 / 2\sigma^2] \quad (19)$$

Using non-informative (10), joint prior, the joint posterior density is:

$$g^*(\mu,\sigma \mid x) \propto (1/\sigma^2)^{(n+2)/2} \exp-[\sum_{i=1}^{n}(x_i - \mu)^2 / 2\sigma^2] \quad (20)$$

$$= (1/\sigma^2)^{(n+2)/2} \exp-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\mu - \bar{x})^2] \quad (21)$$

Where $\bar{x}$ is the sample mean of x_i. The conditional pdf's from the equation above are:

$$g^*(\mu \mid \sigma^2, x) \propto \exp\left\{\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\} \quad (22)$$

$$g^*(\sigma^2 \mid \mu, x) \propto (1/\sigma^2)^{(n+2)/2} \exp-[\sum_{i=1}^{n}(x_i - \mu)^2 / 2\sigma^2] \quad (23)$$

$$g^*(\mu \mid \sigma^2, x) = N(\bar{x}, \sigma^2 / n) \quad (24)$$

The posterior mean is given by:

$$\mu_P = E(x) = \int_{-\infty}^{\infty} g^*(\mu \mid x)\, dx \tag{25}$$

$$\propto \int_{0}^{\infty} x(1/\sigma^2)(1/\sigma^2)^{n/2} \exp-[\sum_{i=1}^{n}(x_i - \mu)^2 / 2\sigma^2]\, dx \tag{26}$$

This is the posterior probability density function of the mean of uncensored data.

CHAPTER 4

EXAMPLES

The following are numerical examples generated from normal populations with mean, μ and standard deviation, σ. The first five examples are generated from N (1, 1) with detection limit, DL=.1.Example 6 is generated from normal population, N (1.306, .134) with two non-detects, Singh and Nocerino (2002). Example 7, Singh and Nocerino (2002), is taken from U.S. EPA RCRA guidance document (1992) with detection limit, 1450, and three non-detects. The posterior probability density function of the mean, $g^*(\mu|\sigma^2,x)$, is plotted using Mathematica and the numerical integration is used to obtain the posterior mean and the upper confidence limit was programmed in Matlab using Simpson quadrature rule with error $<10^{-6}$.

Example 1A (Complete Data Set)

The simulated data set of size 30 was generated from a normal population with mean, $\mu = 1$ and $\sigma = 1$, N (1, 1). The generated data are as follows: .33483, 1.07417, .91798, .53191, 1.58731, 2.72819, .95847, 1.7179, .18525, .43238, 1.35569, 1.95343, .93426, -.46753, -.2097, .33177, 2.63655, -.10443, 2.48921, 3.8581, 1.98537, .01594, 1.01421, .13981, 2.16441, 1.6618, 3.80945, .40988, 1.41659, and 1.22999.The sample mean and the standard deviation using the full uncensored data were 1.27 and 1.099, respectively. The following is the plot of the posterior density of the mean, $g^*(\mu \mid \sigma^2, x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.307.

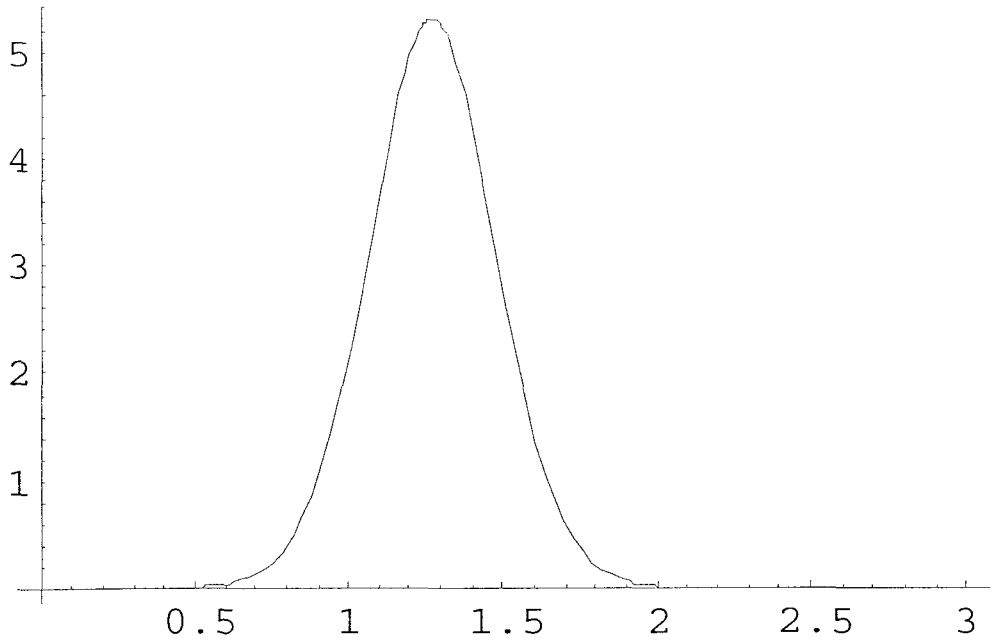

Figure 3: Posterior Density of the mean, $\mu = 1.27$, $\sigma = 1.099$

Example 1B (Censored Data)

The simulated data set of example 1A with DL=.1 and k=4 are the following:

<.1, <.1, <.1, <.1, .33483, 1.07417, .91789, .53191, 1.58731, 2.72819, .95847, 1.7179, .18525, .43238, 1.35569, 1.95343, .93426, .33177, 2.63655, 2.48921, 3.87581, 1.98537, 1.01421, 1.13981, 2.16441, 1.6618, 3.80945, .40988, 1.41659, and 1.22999. The sample mean and the standard deviation obtained using the 26 observed values were: 1.49 and 1.001, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$, using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.5091 and the upper credible limit (UCL) is 1.875.

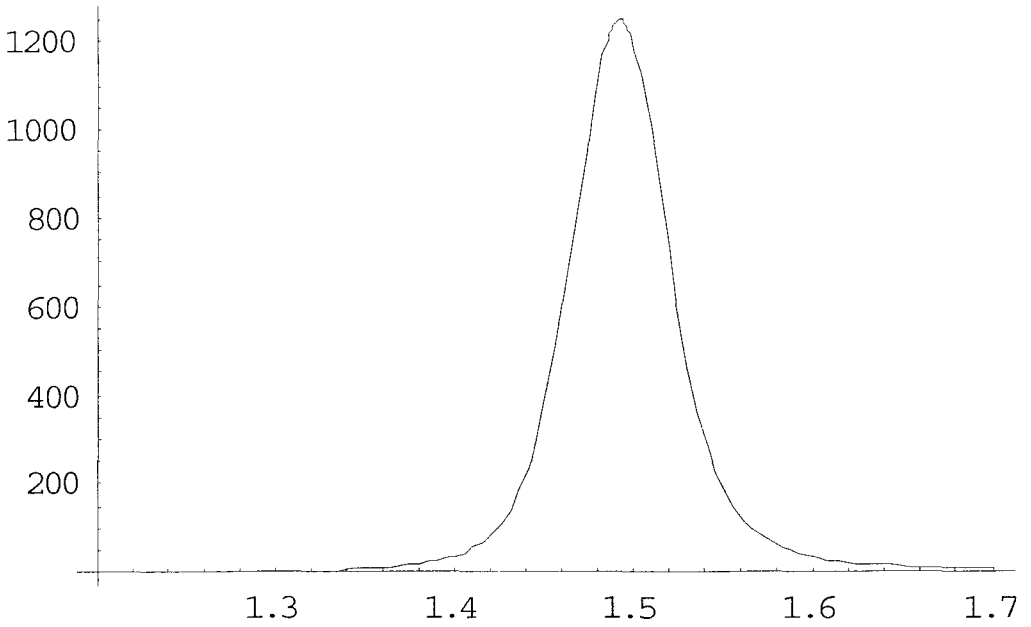

Figure 4: Posterior Density of the mean, $\mu = 1.49$, $\sigma = 1.001$

Example 2A (Full Data)

A simulated data set of size 30 was generated from a normal population with mean, $\mu = 1$ and $\sigma = 1$, N (1, 1), -.34637, 1.23544, .65759, .55156, .73505, 2.30196, .44569, 1.87822, 1.274, .95734, 1.10993, .10149, .89895, 2.13774, 1.35832, 1.30284, 1.99124, .20874, -.64009, -1.57503, 1.51805, 2.0091, 2.60781, -.56341, 1.34461, .88987, .20914, -.4529, 1.76152, and .18125. The sample mean and the standard deviation using the full data were: 0.870 and 0.988, respectively. The following is the plot of the posterior density of the mean, $g^*(\mu \mid \sigma^2, x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 0.967.

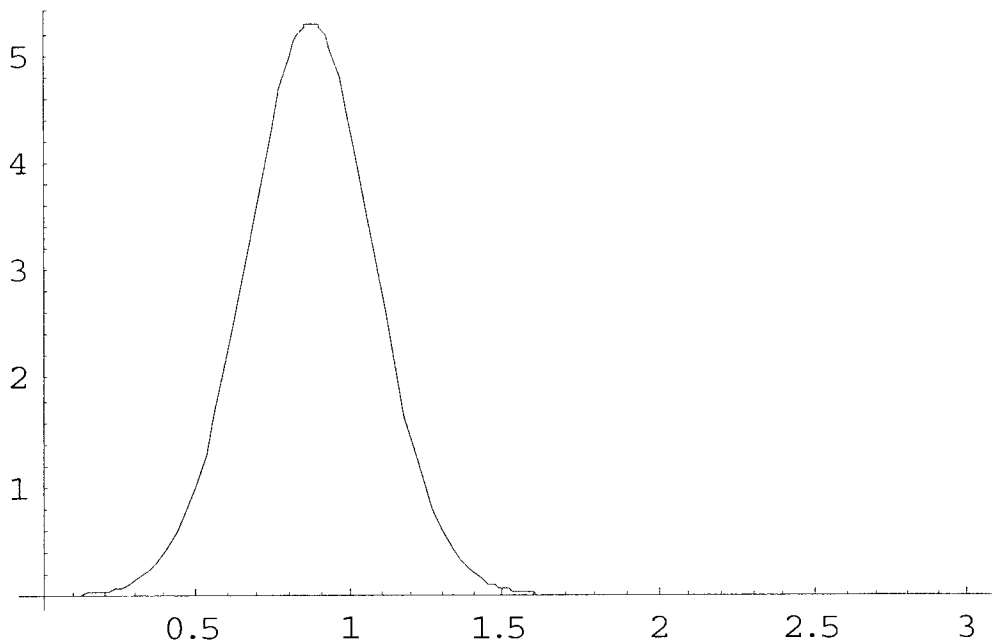

Figure 5: Posterior Density of the mean, $\mu = .87$, $\sigma = .988$

Example 2B (Censored Data)

The simulated data set of example 1A with DL=.1 and k=5 are the following:

<.1, <.1, <.1, <.1, <.1, 1.23544, .65759, .55156, .73505, 2.30196, .44569, 1.87822, 1.274, .95734, 1.10993, .10149, .89895, 2.13774, 1.35832, 1.30284, 1.99124, .20874, 1.51805, 2.0091, 2.60781, 1.34461, .88987, .20914, 1.76152, and .18125. The sample mean and the standard deviation obtained using the 25 observed values were 1.187 and 0.715, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$, using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.2565 and the upper credible limit (UCL) is 1.33.

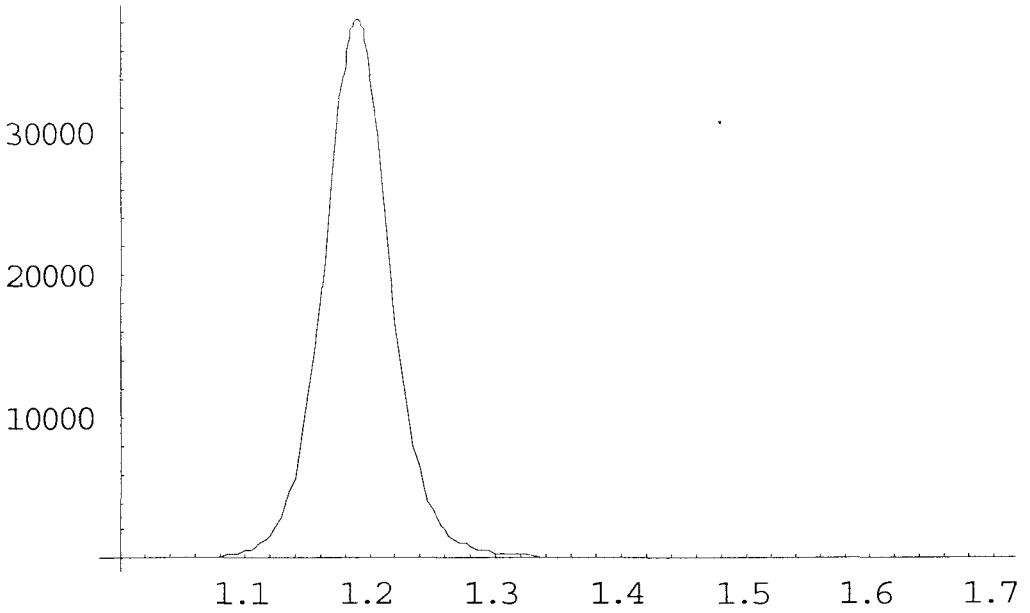

Figure 6: Posterior Density of the mean, $\mu = 1.187$, $\sigma = .715$

Example 3A (Full Data)

A simulated data set of size 30 was generated from normal population with mean, $\mu = 1$ and $\sigma = 1$, N (1, 1).1.1509, 2.33669, 3.03262, 1.41328, 2.12521, 2.18608, 1.91767, 1.42204, 1.78774, -.98267, 2.60349, -.53495, .53363, 1.59111, .8585, .17489, 2.23636, .18885, 1.38405, 1.23115, .54023, 2.40644, .3547, 1.25482, .98104, .63982, 1.56435, 1.33922, 1.03252, and 1.33326. The sample mean and the standard deviation using the full data were 1.27 and 0.921, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2, x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.307.

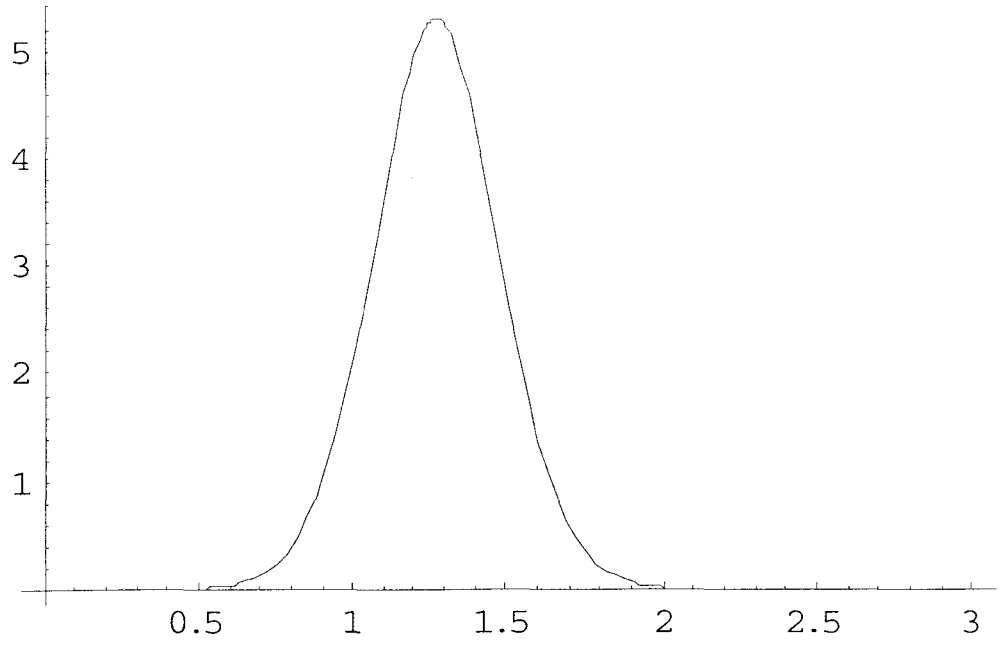

Figure 7: Posterior Density of the mean, $\mu = 1.27$, $\sigma = .921$

Example 3B (Censored Data)

The simulated data set of example 1A with DL=.1 and k=2 are the following:

<.1, <.1, 1.1509, 2.33669, 3.03262, 1.41328, 2.12521, 2.18608, 1.91767, 1.42204, 1.78774, 2.60349, .53363, 1.59111, .8585, .17489, 2.23636, .18885, 1.38405, 1.23115, .54023, 2.40644, .3547, 1.25482, .98104, .63982, 1.56435, 1.33922, 1.03252, and 1.33326. The sample mean and the standard deviation obtained using the 28 observed values were 1.415 and 0.750, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu \mid \sigma^2, x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.5256 and the upper credible limit (UCL) is 1.18.

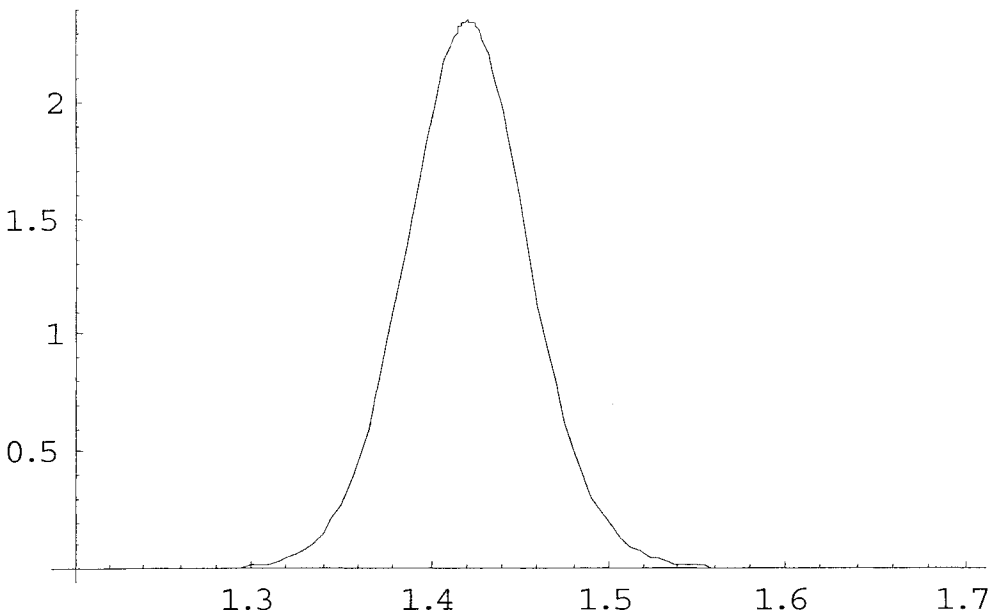

Figure 8: Posterior Density of the mean, $\mu = 1.415$, $\sigma = .750$

Example 4A (Full Data)

A simulated data set of size 30 was generated from normal population with mean, μ=1 and σ=1, N (1, 1), .11126, 1.75206, 1.17052, 3.15024, 3.18094, 1.56179, .927, 2.14169, -.46995, 2.18118, .98145, 1.41042, 3.10198, 2.78779, .71599, .54362, .5441, 2.71058, 2.60982, .77772, 1.80419, -.19731, -1.12471, 1.50846, 1.18456, .50036, .61259, -.95038, 3.26472, and 1.4098. The sample mean and the standard deviation using the full data were 1.33 and 1.216, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.359.

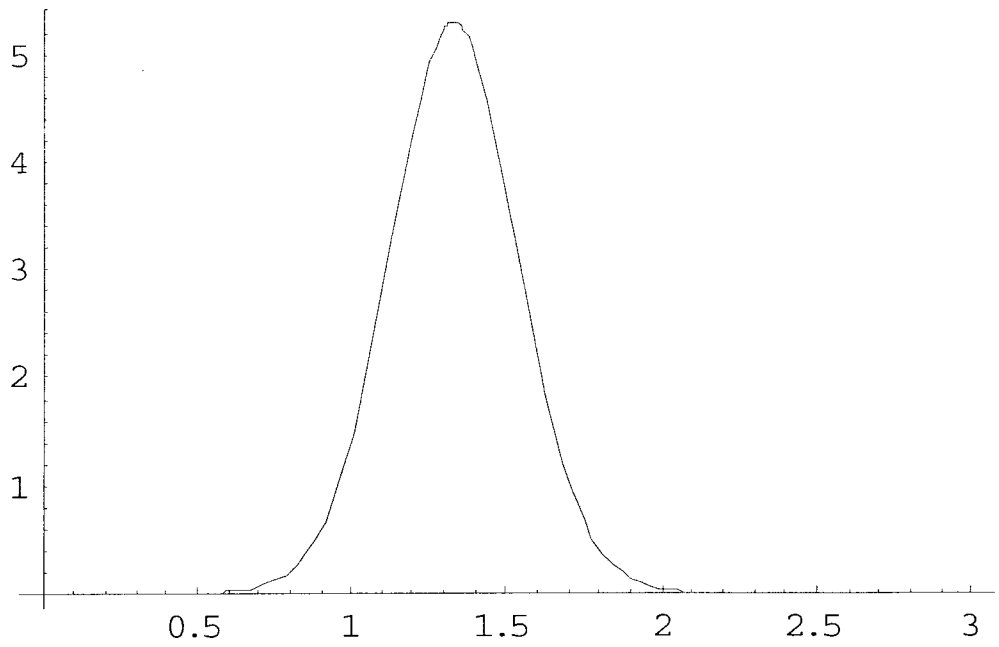

Figure 9: Posterior Density of the mean, μ=1.33, σ=1.216

Example 4B (Censored Data)

The simulated data set of example 1A with DL=.1 and k=4 are the following:

<.1, <.1, <.1, <.1, .11126, 1.75206, 1.17052, 3.15024, 3.18094, 1.56179, .927, 2.14169, 2.18118, .98145, 1.41042, 3.10198, 2.78779, .71599, .54362, .5441, 2.71058, 2.60982, .77772, 1.80419, 1.50846, 1.18456, .50036, .61259, 3.26472, and 1.4098. The sample mean and the standard deviation obtained using the 26 observed values were 1.64 and 0.972, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.7972 and the upper credible limit (UCL) is 1.82.

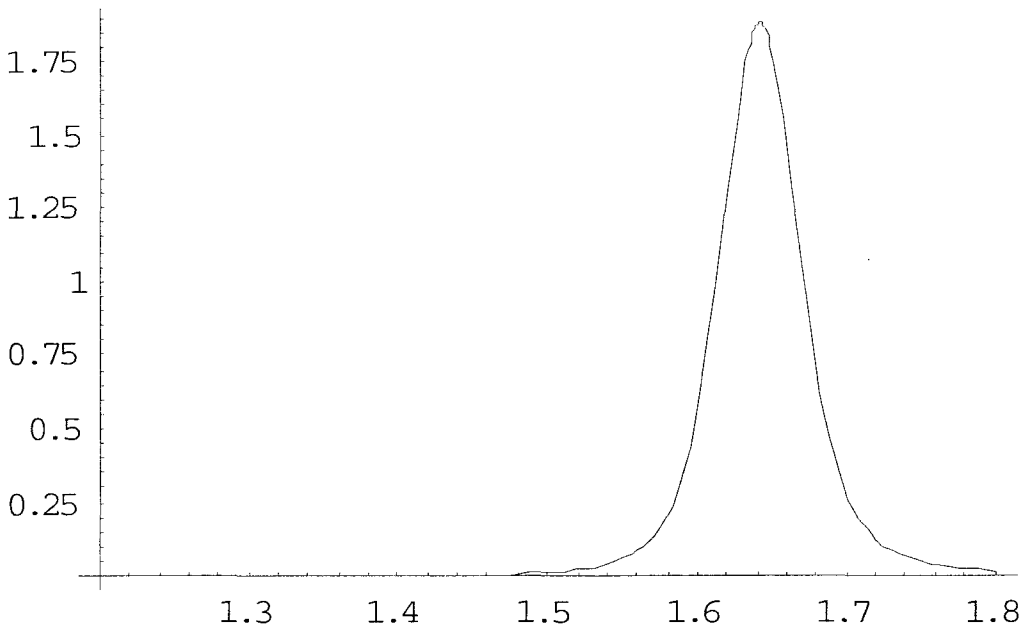

Figure 10: Posterior Density of the mean, $\mu = 1.64$, $\sigma = .972$

Example 5A (Full Data)

A simulated data set of size 30 was generated from normal population with mean, $\mu=1$ and $\sigma=1$, N (1, 1), .98559, 1.72512, .76537, 3.06721, 3.1921, 2.01209, 1.91794, -.11061, .96908, 1.09452, 2.52398, .4659, 1.60952, .54288, -.17748, .74285, -.32055, 1.84595, -.25303, 1.46591, 4.05311, 2.03353, -1.42666, -.13452, -.40622, .88733, 1.35628, 1.36043, 2.10606, and .30362. The sample mean and the standard deviation using the full data were 1.14 and 1.212, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.196.

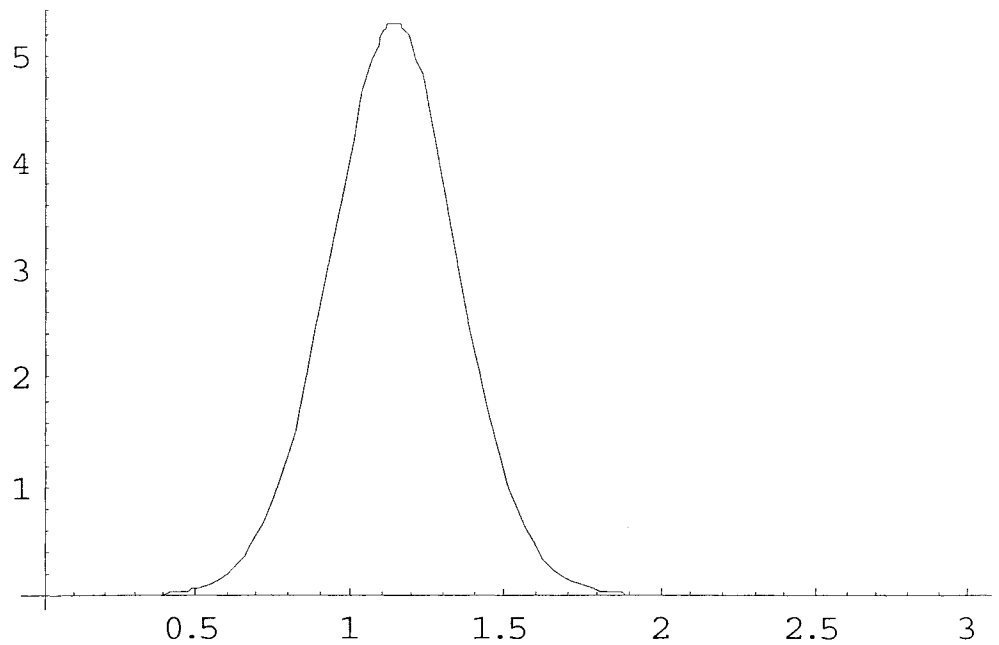

Figure 11: Posterior Density of the mean, $\mu=1.41$, $\sigma=1.212$

Example 5B (Full Data)

The simulated data set of example 1A with DL=.1 and k=7 are the following:

<.1, <.1, <.1, <.1, <.1, <.1, <.1, .98559, 1.72512, .76537, 3.06721, 3.1921, 2.01209, 1.91794, .96908, 1.09452, 2.52398, .4659, 1.60952, .54288, .74285, 1.84595, 1.46591, 1.46591, 4.05311, 2.03353, .88733, 1.35628, 1.36043, 2.10606, and .30362. The sample mean and the standard deviation obtained using the 23 observed values were 1.610 and 0.942, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2,x)$ using Mathematica. The Bayes estimate of the mean μ is the posterior mean 1.6825 and the upper credible limit (UCL) is 1.71.

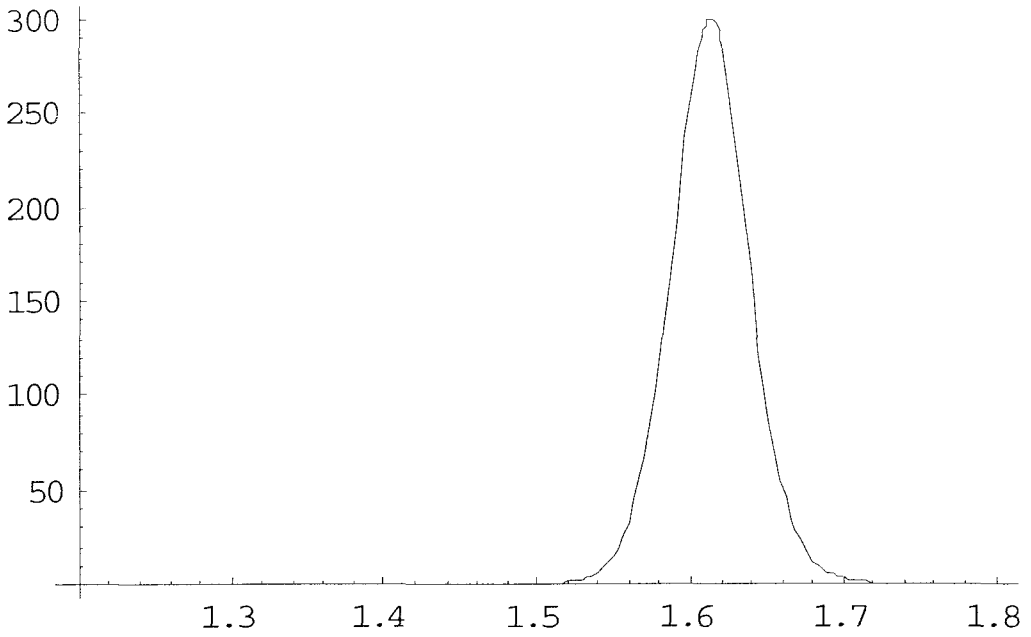

Figure 12: Posterior Density of the mean, $\mu = 1.610$, $\sigma = .942$

Example 6 from Singh and Nocerino (2002)

A simulated data set of size 15 was obtained from a normal population with mean, μ = 1.33 and standard deviation, σ =.2, N (1.33, .2), with detection limit, L=1.0, and k=2. The left-censored data are: <1.0, <1.0, 1.2883, 1.1612, 1.156, 1.3251, 1.1568, 1.5638, 1.2914, 1.3253, 1.2884, 1.4688, 1.4581, 1.3641, and 1.1342. The sample mean and the standard deviation obtained from the 13 observed data values are 1.306 and 0.134, respectively. The following is the posterior probability density plot of the mean: $g^*(\mu|\sigma^2,x)$ of the left-censored data. The Bayes estimate of the mean μ is the posterior mean 1.5123 and the upper credible limit (UCL) is 1.7.

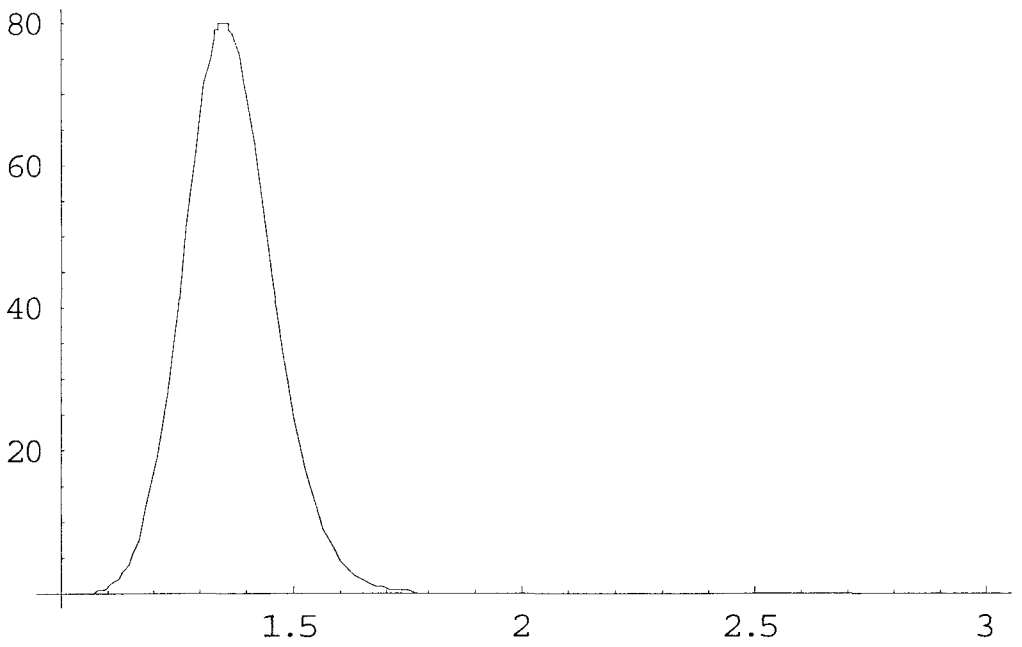

Figure 13: Posterior Density of the mean, μ =1.306, σ =.134

Example 7 from Singh and Nocerino (2002)

This left-censored data set is taken from the U.S. EPA RCRA guidance document [1992]. The detection limit, DL, is 1,450. The data has 3 non-detects and 21 observed values and they are: <1450, <1450, <1450, 1850, 1760, 1710, 1575, 1475, 1780, 1790, 1780, 1790, 1800, 1800, 1840, 1820, 1860, 1780, 1760, 1800, 1900, 1770, 1790, and 1780. The sample mean and standard deviation using the 21 observations are 1771.91 and 92.702, respectively. The following is the posterior probability density plot of the mean: $g^*(\mu|\sigma^2,x)$ of the left-censored data. The Bayes estimate of the mean μ is the posterior mean 1771.7 and the upper credible limit(UCL) is 1775.

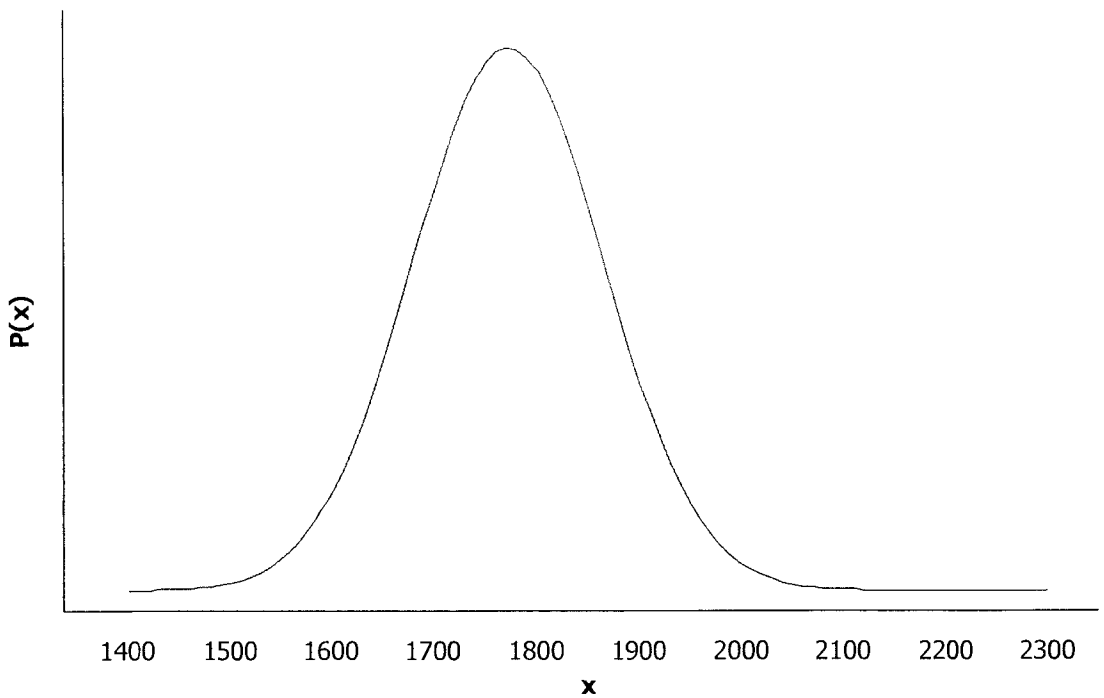

Figure 14: Posterior Density of the mean, $\mu = 1771.9$, $\sigma = 92.702$

The following table is comparison result from Newman, M.C., K.D. Greene, and P.M.Dixon. 1995. Uncensor v4.0. Savannah River Ecology Laboratory. 91 p.

Table 1: Comparison of Different Methods (Uncensored data)

Data Set	Method	Mean	95% Confidence Interval Mean
N=135	Iterative Maximum Likelihood	18.265	17.614-18.961
	Winsorization	18.265	17.613-18.917
	Bayesian Method	19.25	19.5 (UCL)
N=200	Iterative Maximum Likelihood	18.783	18.231-19.335
	Winsorization	18.783	18.230-19.336
	Bayesian Method	19.1	19.7(UCL)

Table 2: Comparison of Different Methods (Censored data)

Data Set	Method	Mean	95% Confidence Interval Mean
N=135 DL=15.67	Iterative Maximum Likelihood	18.269	17.636-18.899
	Winsorization	18.602	17.593-19.261
	Bayesian Method	19.4514	19.9 (UCL)
N=200 DL=15.40	Iterative Maximum Likelihood	18.839	18.31-19.366
	Winsorization	18.868	18.266-19.469
	Bayesian Method	19.2	19.5(UCL)

Table 1. Summary of the Data Simulation and Examples

Examples	Sample Size n	Sample Size Below DL	Detection Limit	Posterior Mean of Uncensored Data	Posterior Mean of Censored Data	95%UCL of Censored Data	95% UCL of uncensored data (bootstrap)	95% UCL of censored data (bootstrap)
Ex1	30	4	.1	1.307	1.5091	1.875	1.6	1.875
Ex2	30	5	.1	.967	1.2565	1.33	1.138	1.429
Ex3	30	2	.1	1.307	1.5256	1.78	1.5	1.418
Ex4	30	4	.1	1.359	1.7972	1.82	1.697	1.933
Ex5	30	7	.1	1.196	1.6825	1.714	1.491	1.947
Ex6	15	2	1	N/A	1.5123	1.70	N/A	1.369
Ex7	24	3	1450	N/A	1771.7	1775	N/A	1797

CHAPTER 5

SUMMARY AND CONCLUSION

This paper is concerned with the Bayesian estimate of the mean of the left-censored data. Considering the non-informative prior, the marginal posterior probability density function of the mean from left-censored data was obtained. However, this density function can not be integrated analytically; numerical integration therefore was implemented to obtain the posterior mean and the upper credible limit. Several numerical examples are presented for illustration.

REFERENCES

Cohen, A.C., Jr., "Estimating the mean and variance of normal populations from singly truncated samples,"Ann. Math.Statist., Vol. 21, pp. 557-569, 1950.

Cohen, A.C., Jr., "Simplified estimators for normal distribution when samples are singly censored or truncated," Technometrics, Vol. 1, No. 3, pp.217-237, 1959.

Cochran, W.G.,"On estimating the mean and the standard deviation of truncated normal distribution ", Jour. Am. Stat. Assn., Vol. 44(1949), pp. 518-525.

Dixon, W.J. and J.W.Tukey. 1968. Approximate behavior of the distribution of Winsorized t (Trimming/ Winsorization 2), Technometrics 10: 83-98.

El-Shaarawi, A.H., "Inferences about the mean from censored water quality data." Water Resources Research, 25, pp. 685-690, 1989.

Gilbert, R. O., "Statistical Methods for Environmental Pollution Monitoring, "Van Nostrand Reinhold, New York, 1987.

Gelit, A., "Estimation for small normal data sets with detection limits," Environmental Science and Technology, 19, pp. 1206-1213, 1985.

Gillion, R.J., and Helsel, D.R., "Estimation of distribution parameters for censored trace level water quality data. 1. Estimation Techniques," water Resources Research, 22, pp.135-146, 1986.

Hald, A. "Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point", Skandinavisk Aktuarietidskrift, Vol. 32 (1949), pp. 119-134.

Jeffreys, H., 1961. Theory of probability (3rd edn.). Oxford University Press, London.

Lee, P.M. 1989. Bayesian Statistics : An Introduction, Oxford University Press, New York.

Newman, M.C., K.D. Greene, and P.M.Dixon. 1995. Uncensor v4.0. Savannah River Ecology Laboratory. 91 p.

Pearson .K. and A.Lee." On the generalized probable error in multiple normal correlation", Biometrika, Vol. 6 (1908), pp. 59-68.

Persson, T., and Rootzen, H., "Simple and highly efficient estimators for a Type I censored normal sample," Boimetrika, 64, pp. 123-128, 1977.

Shumway, A.H., Azari, A.S., Johnson, P. " Estimating mean concentrations under transformation for environmental data with detection limits," Technometrics, Vol. 31, No. 3, pp. 347-356, 1989.

Stevens, W.L. " The truncated normal distribution", appendix to " The calculation of the Time-Mortality Curve" by C.I. Bliss, Annals of Applied Biology, Vol. 24 (1937). Pp. 815-852.

Singh, A., and Nocerinco, J., "Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observation,"

VITA

Graduate College
University of Nevada, Las Vegas

Ahmed Khago

Local Address:
2096 Ramrod Ave #113
Henderson, NV 98014

Degrees:
Bachelor of Science, Mathematics, 1999
University of Wyoming, Laramie

Thesis Title: Bayesian Estimation of the Normal Mean in the Presence of Non-Detects

Thesis Examination Committee:
Chairperson, Dr. Ashok K Singh, Ph.D.
Committee Member, Dr. Rohan Dapatadu, Ph.D.
Committee Member, Dr. Dieudonne, D. Phanord, Ph.D.
Graduate Faculty Representative, Dr. Laxmi, P. Gewali, Ph.D.