Comparison of Different Methods for Estimating Log-normal Means

---

A thesis

presented to

the faculty of the Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Qi Tang

May 2014

Yali Liu Ph.D., Chair

Edith Seier, Ph.D.

Nicole Lewis, Ph.D.

Keywords: Log-normal Distribution, Bayesian Method, Frequentist Method, MSE

ABSTRACT

Comparison of different methods for estimating log-normal means

by

Qi Tang

The log-normal distribution is a popular model in many areas, especially in biostatistics and survival analysis where the data tend to be right skewed. In our research, a total of ten different estimators of log-normal means are compared theoretically. Simulations are done using different values of parameters and sample size. As a result of comparison, "a degree of freedom adjusted" maximum likelihood estimator and Bayesian estimator under quadratic loss are the best when using the mean square error (MSE) as a criterion. The ten estimators are applied to a real dataset, an environmental study from Naval Construction Battalion Center (NCBC), Super Fund Site in Rhode Island.

ACKNOWLEDGMENTS

I would like to thank Dr.Yali Liu, my advisor, for her advice, support and patience during the whole process of my thesis. Special thanks to Dr.Seier and Dr.Lewis, they provided me lots of help with my thesis. I would also like to thank my parents JinQiao Tang and Yan Zhang. Without their support, I couldn't come to ETSU to study. Also I want to express my gratitude to all the people in the Department of Mathematics and Statistics at ETSU who have helped me in the past three years.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

The log-normal distribution is widely used in many areas, such as environmental study, survival analysis, biostatistics and other statistical fields. It is a right skewed distribution with a long tail. Figure 1 displays the log-normal density curves with different parameters. The log-normal distribution has a close association with the normal distribution. By taking the natural logarithm of a random variable, the random variable then will have a normal distribution.

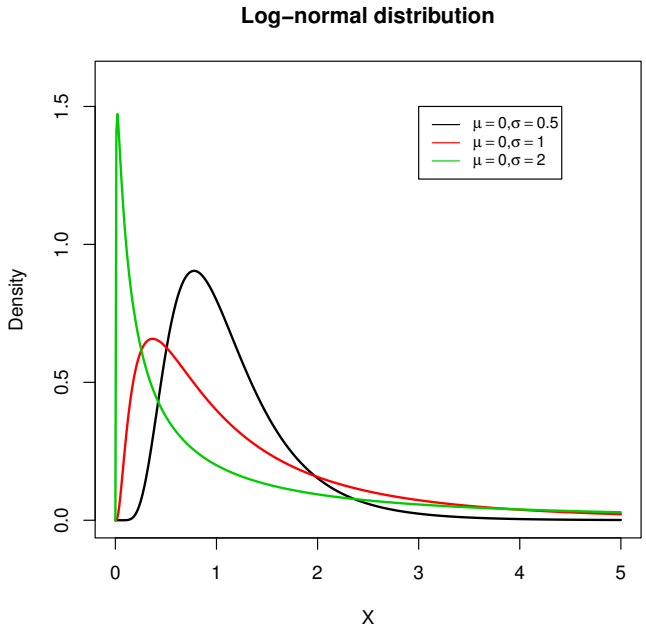

Figure 1: Density curve of several log-normal distributions

One of the important interests is to get the efficient estimator for the log-normal means. There is a long history of people seeking of an estimator of the log-normal means. In 1941, Finney [9] found an unbiased estimator with minimum variance. He also created the Finney function which has been used in other estimators. In 1971, Zellner [6] proposed the minimal mean squared error (MSE) estimator and derived the minimizing value under relative quadratic loss function. Rukhin [5] first provided a generalized form of the log-normal mean estimator in 1986. He also found the generalized prior of the Bayesian estimator. However, the posterior distribution was not provided. In 1998, Zhou [3] developed an estimator based on Zeller's minimal mean squared error (MSE) estimator. In 2008, Longford [1] provided a minimax estimator when the maximum of a parameter is known. In 2012, Enrico and Carlo [2] improved Rukhin's method; they proposed a new prior based on Ruknin's prior, which can be treated as the product of a flat prior.

In this thesis, Chapter 2 discusses eight frequentist methods for estimating log-normal means. For each of the methods, the estimator, the bias and the mean squared error (MSE) are given. Chapter 3 introduces two Bayesian methods for estimating log-normal means. Chapter 4 compares these ten estimators by relative mean squared error and relative bias using graphical displays. Simulations are presented in Chapter 5 to check the theoretical results when real data are involved. A real world example applying these estimators are presented in Chapter 6. These ten estimators are applied to an environmental dataset which has a small sample size. Point estimates for two contaminants are presented. Finally, conclusions are drawn and future work is discussed.

2 FREQUENTISTS METHODS

Let X be the random variable from a log-normal distribution with parameters μ and σ^2. The parameter of common interest is $\theta = e^{p\mu+q\sigma^2}$. Different measures of the distribution have this form and some examples are the median ($p = 1$, q=0), the mode ($p = 1$ and $q = -1$), and the mean ($p = 1$ and $q = 0.5$). In this thesis, only the estimators for the log-normal means are considered.

The mean of X is

$$\mathrm{E}(X) = e^{\mu+\frac{\sigma^2}{2}} \tag{1}$$

and the variance is

$$\mathrm{V}(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \tag{2}$$

Let $\bar{X}$ be the sample mean from the log-normal distribution of size n. The sampling distribution has a mean of

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X) = e^{\mu+\frac{\sigma^2}{2}}, \tag{3}$$

and a variance of

$$\mathrm{V}\left(\bar{X}\right) = \frac{\mathrm{V}(X)}{n} = \frac{(e^{2\mu+\sigma^2})(e^{\sigma^2} - 1)}{n}. \tag{4}$$

Define $Y = \log(X)$. Then Y is normally distributed with mean μ and variance σ^2. Let $\bar{Y}$ be the sample mean of Y. The goal is to find a constant b where $\hat{\theta}(b) = e^{\bar{Y}+bS^2}$ is used to estimate $\theta = e^{\mu+\frac{\sigma^2}{2}}$.

The criterion used to compare different methods are the mean squared error (MSE) and bias. The MSE measures the expected value of the difference between the estimator and the true parameter, i.e., $\mathrm{MSE} = \mathrm{E}(\hat{\theta}(b) - \theta)^2$. The bias is defined as the

difference between the expected value of the estimator and the true parameter, i.e., $\text{Bias}(\hat{\theta}(b)) = \text{E}(\hat{\theta}(b) - \theta)$. To derive these, the expectation and the variance of $e^{\bar{Y}}$ and $e^{2\bar{Y}}$ will be used.

The sample mean, $\bar{Y}$, has a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. Therefore, the exponential of the sample mean, $\bar{Y}$, has a log-normal distribution, i.e.,

$$e^{\bar{Y}} \sim \text{Lognormal}\left(\mu, \frac{\sigma^2}{n}\right).$$

Its expectation and variance are

$$\text{E}\left(e^{\bar{Y}}\right) = e^{\mu + \frac{\sigma^2}{2n}} \tag{5}$$

and

$$\text{V}(e^{\bar{Y}}) = \left(e^{\frac{\sigma^2}{n}} - 1\right) e^{2\mu + \frac{\sigma^2}{n}}. \tag{6}$$

Furthermore, $2\bar{Y}$ has a normal distribution with mean 2μ and variance $\frac{2\sigma^2}{n}$. So we have

$$e^{2\bar{Y}} \sim \text{Lognormal}\left(2\mu, \frac{2\sigma^2}{n}\right).$$

Therefore,

$$\text{E}\left(e^{2\bar{Y}}\right) = e^{2\mu + \frac{2\sigma^2}{n}}. \tag{7}$$

When the expectation and the variance of $e^{\bar{Y}}$ and $e^{2\bar{Y}}$ are obtained, we can use them to derive the bias and MSE. In the following sections, different estimation methods will be considered.

2.1 The Naive Estimator

In statistics, a natural estimator of a distribution mean is the sample mean. In this thesis, it is called the naive estimator. Therefore, the naive estimator is

$$\hat{\theta}_1 = \bar{X}.$$

This estimator is unbiased, thus

$$\text{Bias}\left(\hat{\theta}_1\right) = 0. \tag{1}$$

And the MSE is equal to $\text{V}(\bar{X})$, i.e.,

$$\text{MSE}\left(\hat{\theta}_1\right) = \text{V}\left(\hat{\theta}_1\right) = \text{V}\left(\bar{X}\right) = \frac{\left(e^{\sigma^2} - 1\right)\left(e^{2\mu+\sigma^2}\right)}{n}. \tag{2}$$

The naive estimator is easy to calculate and it is unbiased. However, this estimator can be inefficient when σ^2 is large and sample size is small.

2.2 The Maximum Likelihood Estimator (MLE)

Another commonly used estimator is the maximum likelihood estimator (MLE). The basic idea in this section is to find the MLE of μ and σ^2 and ultimately θ.

The following terminology is defined for future use. The sample variance of Y is

$$S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}. \tag{8}$$

The MLE for σ^2 is

$$\hat{\sigma^2} = \frac{\sum(Y_i - \bar{Y})^2}{n} = \frac{(n-1)S^2}{n}. \tag{9}$$

The maximum likelihood estimator for μ is $\bar{X}$. Therefore the maximum likelihood estimator for θ is

$$\hat{\theta}_2 = e^{\bar{Y} + \frac{(n-1)S^2}{2n}}.$$

The random variable $V = \frac{(n-1)S^2}{\sigma^2}$ has a Chi-square distribution with $k = n-1$ degrees of freedom, i.e., $V \sim \chi^2_{n-1}$.

The Chi-square density function with $k = n-1$ degrees of freedom is

$$f(t) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} t^{\frac{k}{2}-1} e^{-\frac{t}{2}}.$$

To find the MSE of $\hat{\theta}_2$, we need to derive $\mathrm{E}(\hat{\theta}_2)$ and $\mathrm{E}(\hat{\theta}_2^2)$. First we have

$$\begin{aligned}
\mathrm{E}(\hat{\theta}_2) &= \mathrm{E}\left(e^{\bar{Y}+\frac{(n-1)S^2}{2n}}\right) = \mathrm{E}\left(e^{\bar{Y}}\right)\mathrm{E}\left(e^{\frac{(n-1)S^2}{2n}}\right) \\
&= e^{\mu+\frac{\sigma^2}{2n}} \int_0^\infty e^{t\frac{\sigma^2}{2n}} \frac{t^{\frac{n-1}{2}-1}e^{-\frac{t}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} dt \\
\overset{v=t\left(1-\frac{\sigma^2}{n}\right)}{=} & e^{\mu+\frac{\sigma^2}{2n}} \int_0^\infty \frac{v^{\frac{n-1}{2}-1}e^{-\frac{v}{2}}}{\left(1-\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}-1} 2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \frac{dv}{1-\frac{\sigma^2}{n}} \\
&= e^{\mu+\frac{\sigma^2}{2n}} \frac{1}{\left(1-\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}}} \\
&= e^{\mu+\frac{\sigma^2}{2n}} \left(\frac{n}{n-\sigma^2}\right)^{\frac{n-1}{2}}
\end{aligned}$$

Using equation (7), $\mathrm{E}(\hat{\theta}_2^2)$ is obtained using a similar process as we used to find $\mathrm{E}\left(\hat{\theta}_2\right)$. That is,

$$\begin{aligned}
\mathrm{E}(\hat{\theta}_2^2) &= \mathrm{E}\left(e^{2\bar{Y}+\frac{(n-1)S^2}{2n}}\right) = \mathrm{E}\left(e^{2\bar{Y}}\right)\mathrm{E}\left(e^{\frac{(n-1)S^2}{2n}}\right) \\
&= e^{2\mu+\frac{2\sigma^2}{n}} \int_0^\infty e^{t\frac{\sigma^2}{n}} \frac{t^{\frac{n-1}{2}-1}e^{\frac{-t}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} dt \\
\overset{v=t\left(1-\frac{\sigma^2}{n}\right)}{=} & e^{2\mu+\frac{2\sigma^2}{n}} \int_0^\infty \frac{v^{\frac{n-1}{2}-1}e^{-\frac{v}{2}}}{\left(1-\frac{2\sigma^2}{n}\right)^{\frac{n-1}{2}-1} 2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \frac{dv}{1-\frac{2\sigma^2}{n}} \\
&= e^{2\mu+\frac{2\sigma^2}{n}} \frac{1}{\left(1-\frac{2\sigma^2}{n}\right)^{\frac{n-1}{2}}} \\
&= e^{2\mu+\frac{2\sigma^2}{n}} \left(\frac{n}{n-2\sigma^2}\right)^{\frac{n-1}{2}}
\end{aligned}$$

Therefore Bias($\hat{\theta}_2$) and MSE($\hat{\theta}_2$) are obtained as

$$\text{Bias}\left(\hat{\theta}_2\right) = e^{\mu+\frac{\sigma^2}{n}}\left(\frac{n}{n-\sigma^2}\right)^{\frac{n-1}{2}} - e^{\mu+\frac{\sigma^2}{2}} \quad (3)$$

and

$$\begin{aligned}\text{MSE}\left(\hat{\theta}_2\right) &= \text{E}\left(\hat{\theta}_2-\theta_2\right)^2 = \text{E}\left(\hat{\theta}_2^2\right) - 2\theta\text{E}\left(\hat{\theta}_2\right) + \theta_2^2 \\ &= e^{2\mu+\sigma^2}\left(e^{\sigma^2(\frac{2}{n}-1)}\left(1-\frac{2\sigma^2}{n}\right)^{\frac{-(n-1)}{2}} - 2e^{\frac{\sigma^2}{2}(\frac{1}{n}-1)}\left(1-\frac{\sigma^2}{n}\right)^{\frac{-(n-1)}{2}} + 1\right). \quad (4)\end{aligned}$$

Similar to the naive estimator, the maximum likelihood estimator is also easy to carry out, thus making it convenient to use in practice. However this estimator tends to be overestimated when sample size is small and is inefficient for large values of σ^2.

2.3 Approximately Minimum Mean Squared Error Estimator

Recall that MLE estimator has the form of $\hat{\theta} = e^{\bar{Y}+bS^2}$, where b is a constant. Longford [1] proposed a method to find the value of b which can make the MSE as small as possible. Thus, he provided this approximately minimum MSE estimator. The main idea is to solve b of the the minimize MSE.

Define the estimator as $\hat{\theta}_3 = e^{\bar{Y}+bS^2}$.

First, we find the $\mathrm{E}\left(\hat{\theta}_3\right)$ to obtain the value of MSE.

$$\begin{aligned} E(\hat{\theta}_3) &= \mathrm{E}\left(e^{\bar{Y}+bS^2}\right) = \mathrm{E}\left(e^{\bar{Y}}\right)\mathrm{E}(e^{bS^2}) \\ &= e^{\mu+\frac{\sigma^2}{2n}}\left(\frac{n-1}{n-1-2b\sigma^2}\right)^{(n-1)/2}. \end{aligned} \tag{10}$$

The MSE is

$$\mathrm{MSE}\left(\hat{\theta}_3\right) = e^{2\mu}\left(e^{\frac{2\sigma^2}{n}}\left(\frac{n-1}{n-1-4b\sigma^2}\right)^{(n-1)/2} - 2e^{\frac{\sigma^2}{2n}+\frac{\sigma^2}{2}}\left(\frac{n-1}{n-1-2b\sigma^2}\right) + e^{\sigma^2}\right). \tag{11}$$

To get the minimum of the MSE, the derivative of MSE with respect to b needs to be taken and so we have

$$\frac{\partial\left(MSE(\hat{\theta}_3)\right)}{\partial b} = 2\sigma^2 e^{2\mu}\left(e^{\frac{2\sigma^2}{n}}\left(\frac{n-1}{n-1-4b\sigma^2}\right)^{(n-1)/2+1} - e^{\frac{\sigma^2}{2n}+\frac{\sigma^2}{2}}\left(\frac{n-1}{n-1-4b\sigma^2}\right)^{(n-1)/2+1}\right).$$

Let $\frac{\partial\left(MSE(\hat{\theta}_3)\right)}{\partial b} = 0$. This implies that:

$$e^{\frac{2\sigma^2}{n}}\left(\frac{n-1}{n-1-4b\sigma^2}\right)^{(n-1)/2+1} = e^{\frac{\sigma^2}{2n}+\frac{\sigma^2}{2}}\left(\frac{n-1}{n-1-2b\sigma^2}\right)^{(n-1)/2+1}.$$

Taking the natural logarithm on both sides, we obtain

$$\frac{2\sigma^2}{n} + \left(\frac{n-1}{2}+1\right)\ln\frac{n-1}{n-1-4b\sigma^2} = \frac{\sigma^2}{2n} + \frac{\sigma^2}{2} + \left(\frac{n-1}{2}+1\right)\ln\frac{n-1}{n-1-2b\sigma^2}.$$

Combine the similar terms of the equation. It turns out to be

$$\frac{n-1-2b\sigma^2}{n-1-4b\sigma^2} = \exp\left(\frac{2\sigma^2}{(n-1)+2}\left(\frac{1}{2}-\frac{3}{2n}\right)\right).$$

Let $D_a = \exp\left(\frac{2\sigma^2}{(n-1)+2}\left(\frac{1}{2}-\frac{3}{2n}\right)\right)$. The constant b is equal to

$$b = \frac{n-1}{2\sigma^2}\frac{D_a-1}{2D_a-1}.$$

Substituting the constant b into the formula (10) and (11), the bias and the MSE are

$$\text{Bias}(\hat{\theta}_3) = e^{\mu+\frac{\sigma^2}{2n}}\left(\frac{n-1}{n-1-2b\sigma^2}\right)^{(n-1)/2} - e^{\mu+\frac{\sigma^2}{2}} \quad (5)$$

and

$$\text{MSE}(\hat{\theta}_3) = e^{2\mu}\left(e^{\sigma^2} - 2e^{\frac{\sigma^2}{2}+\frac{\sigma^2}{2n}}\left(1-\frac{2b\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} + e^{\frac{2\sigma^2}{n}}\left(1-\frac{4b\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}\right). \quad (6)$$

The "approximately minimum MSE estimator" is easy to calculate and implement. It is efficient for both small and large values of σ^2. It will be used as the reference for comparisons of different methods.

2.4 Approximately Unbiased Estimator

The previous section uses the minimum MSE to find the constant b. Different ways to find b have been proposed. Longford [1] proposed an approximately unbiased estimator. This method uses the unbiased estimating equation to obtain the value of b. The estimator has the form

$$\hat{\theta}_4 = e^{\bar{Y}+bS^2}.$$

If $\hat{\theta}_4$ is unbiased for θ, i.e., $\mathrm{E}(\hat{\theta}) = \theta$, then

$$e^{\mu+\frac{\sigma^2}{2n}} \left(\frac{n-1}{(n-1)-2b\sigma^2} \right)^{(n-1)/2} = e^{\mu+\frac{\sigma^2}{2}},$$

$$\frac{n-1-2b\sigma^2}{n-1} = \exp\left[\frac{-2\sigma^2}{n-1} \left(\frac{1}{2} - \frac{1}{2n} \right) \right].$$

Therefore solving for the constant b, we have

$$b = \frac{n-1}{2\sigma^2} \left[1 - \exp\left(-\frac{2\sigma^2}{n-1} \left(\frac{1}{2} - \frac{1}{2n} \right) \right) \right].$$

The bias and the MSE of this estimator are

$$\mathrm{Bias}\left(\hat{\theta}_4\right) = e^{\mu+\frac{\sigma^2}{2n}} \left(\frac{n-1}{n-1-2b\sigma^2} \right)^{(n-1)/2} - e^{\mu+\frac{\sigma^2}{2}} \qquad (7)$$

and

$$\mathrm{MSE}\left(\hat{\theta}_4\right) = e^{2\mu} \left(e^{\sigma^2} - 2e^{\frac{\sigma^2}{2}+\frac{\sigma^2}{2n}} \left(1 - \frac{2b\sigma^2}{n-1} \right)^{-\frac{n-1}{2}} + e^{\frac{2\sigma^2}{n}} \left(1 - \frac{4b\sigma^2}{n-1} \right)^{-\frac{n-1}{2}} \right). \qquad (8)$$

The approximately unbiased estimator has a simple form and it is an unbiased estimator, making it is easier to be applied. However, when the sample size is small, it returns a large MSE. As σ^2 gets large, the estimator becomes inadequate.

2.5 Minimax Estimator

In addition to the previous two estimators, Longford [1] proposed a minimax estimator with the same form. The basic idea for this method is using a minimax method to find b.

When $0 < b < \frac{1}{4}\frac{n-1}{\sigma^2}$, Longford [1] proved that the variance of the estimator is an increasing function of σ^2. Therefore, the MSE is also an increasing function of σ^2. He drew a conclusion that there must be a specified value σ^2_{mx} which is the upper bound of σ^2. When $\sigma^2_{mx} = \sigma^2$, the constant b solving from the equation can make the estimator efficient.

The estimator is

$$\hat{\theta}_5 = e^{\bar{Y}+bS^2},$$

where $b = \frac{n-1}{2\sigma^2_{mx}}\frac{D_{a,mx}-1}{2D_{a,mx}-1}$ and $D_{a,mx} = \exp\left(\frac{2\sigma^2_{mx}}{(n-1)+2}\left(\frac{1}{2}-\frac{3}{2n}\right)\right)$.
The bias and the MSE are

$$\text{Bias}(\hat{\theta}_5) = e^{\mu+\frac{\sigma^2}{2n}}\left(\frac{n-1}{n-1-2b\sigma^2}\right)^{(n-1)/2} - e^{\mu+\frac{\sigma^2}{2}} \qquad (9)$$

and

$$\text{MSE}(\hat{\theta}_5) = e^{2\mu}\left(e^{\sigma^2} - 2e^{\frac{\sigma^2}{2}+\frac{\sigma^2}{2n}}\left(1-\frac{2b\sigma^2}{n-1}\right)^{-\frac{n-1}{2}} + e^{\frac{2\sigma^2}{n}}\left(1-\frac{4b\sigma^2}{n-1}\right)^{-\frac{n-1}{2}}\right). \qquad (10)$$

Unlike the other estimators, the minimax estimator requires the maximum of parameter σ^2 to be known in advance. In certain real analysis situations, the estimator can be hard to handle. However, it performs very well for both the small and large sample sizes regardless of the size of σ^2.

2.6 The Uniformly Minimum Variance Unbiased Estimator (UMVUE)

In this section we summarize an unbiased estimator called the uniformly minimum variance unbiased estimator (UMVUE). It was proposed by Finney [9]. Recall that the parameter of interest is $\theta = e^{\mu+\frac{\sigma^2}{2}}$. The idea is to seek an unbiased estimator which is a function of $\bar{Y}$ and S^2. As $\mathrm{E}(e^{\bar{Y}}) = e^{\mu+\frac{\sigma^2}{2n}}$, we need to find a function of S^2 which is unbiased for $e^{\frac{\sigma^2}{2}-\frac{\sigma^2}{2n}}$.

The estimator is

$$\hat{\theta}_6 = e^{\bar{Y}} g\left(\frac{(n-1)S^2}{2}\right),$$

where $g(t)$ is the Finney's function with the expression of

$$g(t) = \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n-1}{2}\right)}{i!\Gamma(\frac{n-1}{2}+i)} \left(\frac{n-1}{2n}t\right)^i.$$

It can be shown that the expectation of $\hat{\theta}_6$ is

$$\mathrm{E}(\hat{\theta}_6) = \mathrm{E}\left(e^{\bar{Y}}\right) \mathrm{E}\left(g\left(\frac{(n-1)S^2}{2}\right)\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = 0.$$

which is θ. Therefore, the bias and the MSE for $\hat{\theta}_6$ are

$$\mathrm{Bias}(\hat{\theta}_6) = \mathrm{E}(\hat{\theta}_6) - \theta = 0 \qquad (11)$$

and

$$\mathrm{MSE}(\hat{\theta}_6) = e^{2\mu+\sigma^2}\left(e^{\frac{\sigma^2}{n}} g\left(\frac{(n-1)\sigma^4}{2n}\right) - 1\right). \qquad (12)$$

Since the UMVU estimator includes Finney's function, it is difficult to calculate in reality. The estimator is inefficient when the sample size is small. On the other hand, this is an unbiased estimator and it works well as the sample size gets large.

2.7 A Conditional Mean Squared Error Estimator

In this section, the estimator discussed has the same form as the UMVU estimator. Zellner [6] noted that the class of estimators is $Ce^{\bar{Y}}$ where C is a constant. The conditional minimum MSE estimator for θ given σ^2 is $\theta = \exp\left(\bar{Y} + \frac{(n-3)\sigma^2}{2n}\right)$.

Since Zellner's estimator has an unknown parameter σ^2, Zhou [3] proposed a new estimator called a conditional MSE estimator. Because $\mathrm{E}(e^{\bar{Y}}) = \exp\left(\mu + \frac{\sigma^2}{2n}\right)$, an unbiased estimator is needed for $\exp\left(\frac{(n-3)\sigma^2}{2n} - \frac{\sigma^2}{2n}\right)$, and it turns out to be $g\left(\frac{(n-4)S^2}{2}\right)$. Therefore,the estimator is

$$\hat{\theta}_7 = \exp\left(\bar{Y}\right) g\left(\frac{n-4}{2}S^2\right),$$

where $g(t)$ is the Finney's function. It follows that

$$\mathrm{E}(\hat{\theta}_7) = \mathrm{E}\left(e^{\bar{Y}}\right)\mathrm{E}\left(g\left(\frac{(n-4)S^2}{2}\right)\right) = e^{\mu+\frac{\sigma^2}{2}-\frac{3\sigma^2}{2n}}.$$

The bias and the MSE are

$$\mathrm{Bias}(\hat{\theta}_7) = \mathrm{E}(\hat{\theta}_7) - \theta = e^{\mu+\frac{\sigma^2}{2}}\left(e^{\frac{-3\sigma^2}{2n}} - 1\right) \tag{13}$$

and

$$\mathrm{MSE}(\hat{\theta}_7) = e^{2\mu+\sigma^2}\left(e^{\frac{-2\sigma^2}{n}} g\left(\frac{(n-4)^2\sigma^4}{2n(n-1)}\right) - 2e^{\frac{-3\sigma^2}{2n}} + 1\right). \tag{14}$$

Because a conditional MSE estimator involves Finney's function, it is inconvenient for application. This estimator tends to underestimate the true value of the log-normal mean. Yet, it is very efficient for the small sample size.

2.8 "A Degree of Freedom Adjusted" Maximum Likelihood Estimator

Shen [4] proposed another method to estimate the log-normal mean. He used the second-order asymptotic to find a constant b which can minimize the MSE. Let $b = \frac{1}{d+n}$, where $d + n > 0$, both d and n are constants. The term b is then expanded to $\frac{1}{n} - \frac{d}{n^2} + o(\frac{1}{n^2})$. Recall that

$$\text{MSE} = e^{2\mu+\sigma^2}\left[e^{\frac{2-n}{2n}\sigma^2}(1-2c\sigma^2)^{\frac{-(n-1)}{2}} - 2e^{\frac{1-n}{2n}\sigma^2}(1-c\sigma^2)^{\frac{-(n-1)}{2}} + 1\right].$$

Let $W = \frac{\text{MSE}}{e^{2\mu+\sigma^2}}$. then minimizing MSE is the same as minimizing W. Substituting b into W and using Taylor's expansion,

$$W = \frac{\sigma^2}{n}\left[1 + \frac{\sigma^2}{2} + \frac{\sigma^2}{4n}(d^2 - (8+3\sigma^2)d + 8\sigma^2 + \frac{7}{4}\sigma^4)\right] + o(\frac{1}{n^2}).$$

Thus, we only need to minimize $d^2-(8+3\sigma^2)d$ part. Because this is a quadratic form, when $d = 4 + \frac{3}{2}\sigma^2$, W reaches the minimal value. Therefore $b = \frac{1}{n+4+\frac{3\sigma^2}{2}}$. Replacing σ^2 by S^2, the estimator is

$$\hat{\theta}_8 = \exp\left(\bar{Y} + \frac{(n-1)S^2}{2(n+4)+3S^2}\right).$$

The expectation of $\hat{\theta}_8$ is

$$\begin{aligned}\text{E}(\hat{\theta}_8) &= \text{E}(e^{\bar{Y}})\text{E}\left(e^{\frac{(n-1)S^2}{2(n+4)+3S^2}}\right),\\ &= e^{\mu+\frac{\sigma^2}{2n}}\int_0^\infty e^{\frac{(n-1)S^2}{2(n+4)+3S^2}}ds,\end{aligned}$$

Let $V = (n-1)\frac{S^2}{\sigma^2}$. Then $V \sim \chi^2_{n-1}$ and

$$\text{E}(\hat{\theta}_8) = e^{\mu+\frac{\sigma^2}{2n}}\int_0^\infty e^{\frac{(n-1)V}{\frac{2(n+4)(n-1)}{\sigma^2}+3V}}f(V)dV.$$

The bias and the MSE are

$$\text{Bias} = \text{E}(\hat{\theta}_8) - \theta \tag{15}$$

and

$$\text{MSE}(\hat{\theta}_8) = \exp(2\mu + \sigma^2)\left(e^{[(2-n)/n]\sigma^2} f_1 - 2e^{[(1-n)/2n]\sigma^2} f_2 + 1\right), \tag{16}$$

where

$$f_1 = \text{E}\left[\exp\left(\frac{2(n-1)S^2}{2(n+4)+3S^2}\right)\right]$$
$$f_2 = \text{E}\left[\exp\left(\frac{(n-1)S^2}{2(n+4)+3S^2}\right)\right].$$

Although a degree of freedom adjusted MLE tends to underestimate the true value, it has a very small MSE when σ^2 is small or moderate. It also performs well when the sample size gets large. This estimator is recommended when MSE is used as the criterion for comparison.

3 BAYESIAN METHODS

The previous chapter discussed eight different frequentist methods. In this chapter, Bayesian methods will be introduced. In Bayesian statistics, the prior distribution is about what we already know for the parameters before the data is collected. The likelihood function is the probability of the observed data given the known parameters. The posterior distribution is the probability of the parameter given the observed data. Bayes theorem expresses the relationships among the prior probability, the likelihood function, and the posterior probability. The formula can be expressed as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

For the Bayesian method in log-normal distributions, Rukhin [5] proposed a generalized prior,

$$p(\sigma) \propto \sigma^{-2\nu+n-2} \exp\left[-\sigma^2(\frac{\gamma^2}{2} - (1 - \frac{2}{n}))\right],$$

where ν and γ are prior parameters. The generalized form of the estimator is $\delta = e^{\bar{Y}} g\left((n-1)S^2\right)$. Enrico and Carlo [2] proposed a new prior based on Rukhin's prior, this can be seen as the product of a flat prior. The following two sections will discuss the methods proposed by Enrico and Carlo [2]. They are Bayes estimator under quadratic loss and under relative quadratic loss, respectively.

3.1 Bayes Estimator Under Quadratic Loss

For the Bayesian estimators, it is important to know the prior and to derive the posterior distribution. The prior proposed by Enrico and Carlo [2] is

$$p(\sigma^2) \propto (\sigma^2)^{-\nu+\frac{n}{2}-3/2} \exp\left[-\sigma^2(\frac{\psi^2}{2} - 2(b - a^2/n))\right],$$

where $a = 1$, $b = \frac{1}{2}$, ν and ψ are prior parameters. Define $\lambda = -\nu + n/2 - 1/2$, $\gamma^2 = \psi^2/2 - 2(b - a^2/n)$. The prior is a limit of a generalized inverse gamma distribution, $\mathrm{GIG}(\lambda, \delta, \gamma)$ as $\delta \to 0$.

Define $\eta = \log(\theta)$. The distribution of η based on the prior is

$$\eta \sim \mathrm{GH}(\bar{\lambda}, \bar{\alpha}, \bar{\beta}, \bar{\delta}, \bar{\mu}),$$

where $\bar{\lambda} = \lambda - \frac{n-1}{2}, \bar{\alpha} = \sqrt{n(\gamma^2 + \frac{n}{4})}, \bar{\beta} = \frac{n}{2}, \bar{\delta} = \sqrt{\frac{1}{n}((n-1)S^2 + \delta^2)}, \bar{\mu} = \bar{Y}$. This is a generalized hyperbolic (GH) distribution.

The density function of the GH is,

$$f(x) = \frac{(\frac{\gamma}{\delta})^\lambda}{\sqrt{2\pi}K_\lambda(\delta_\gamma)} \frac{K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right)}{\sqrt{\delta^2 + (x-\mu)^2}\alpha} \exp(\beta(x-\mu)),$$

where K is the Bessel-K function.

The posterior distribution of $\theta|x$ is a $\mathrm{Log} - \mathrm{GH}$ distribution.

The moment generating function is used to find the expectation and the variance of the estimator. The results are

$$\mathrm{E}(\theta|x) = \mathrm{M}_{\mathrm{GH}}(1) \tag{12}$$

and

$$\mathrm{V}(\theta|x) = \mathrm{M}_{\mathrm{GH}}(2) - [\mathrm{M}_{\mathrm{GH}}(1)]^2. \tag{13}$$

The quadratic loss function is

$$L(\hat{\theta}, \theta) = [\theta - \hat{\theta}]^2 = [e^{\mu + \frac{\sigma^2}{2}} - \hat{\theta}]^2.$$

It follows that

$$\mathrm{E}(L|\sigma^2) = \mathrm{E}(\theta^2) - 2\hat{\theta}\mathrm{E}(\theta) + \hat{\theta^2}. \tag{14}$$

The minimizing value for $\hat{\theta}$ is obtained by

$$0 = 0 - 2\mathrm{E}(\theta) + 2\hat{\theta}.$$

Therefore the expression for the Bayes estimator under quadratic loss is

$$\hat{\theta} = \mathrm{E}(\theta).$$

Based on formula (12) and (13), the formula for $\hat{\theta}$ is

$$\begin{aligned} \hat{\theta} &= \mathrm{E}(\theta|x) = \mathrm{M}_{\mathrm{GH}}(1) \\ &= e^{a\bar{Y}} \left(\frac{\gamma^2}{\gamma^2 - (\frac{a^2}{n} + 2b)} \right)^{(\lambda - \frac{n-1}{2})/2} \times \frac{K_{\lambda - \frac{n-1}{2}} \sqrt{(\gamma^2 - \frac{a^2}{n} - 2b)((n-1)S^2 + \delta^2)}}{K_{\lambda - \frac{n-1}{2}} \sqrt{((n-1)S^2 + \delta^2)\gamma^2}}. \end{aligned}$$

Since the Bessel function is difficult to calculate, the author used a small argument approximation to replace the Bessel K function and got

$$\hat{\theta} \approx \exp(a\bar{Y}) \exp\left[\frac{-((n-1)S^2 + \delta^2)(a^2 + 2nb)}{4n(\lambda - \frac{n-3}{2})} \right].$$

The λ value is obtained by minimizing MSE, which is

$$\lambda = \frac{n-3}{2} - \frac{(n-1)(a^2 + 2nb)}{4nc} - \frac{(a^2 + 2nb)}{4nc} \frac{\delta^2}{\sigma^2},$$

where $c = b - 3a^2/2n$. Plugging λ in the approximate formula, letting $\delta = \sigma^2$ and neglecting σ^4, then $\hat{\theta} = \exp\left(\bar{Y} + \frac{S^2(n-3)(n-1)}{2n(n-1) + 2n\sigma^2}\right)$, which is close to the estimator proposed by Shen [4]. Thus, there is some relationship between Bayes estimator and non-Bayes estimator.

Replacing σ^2 by its unbiased estimator S^2, the estimator is

$$\hat{\theta}_9 = \exp\left(\bar{Y} + \frac{S^2(n-3)(n-1)}{2n(n-1)+2nS^2}\right).$$

The expectation of $\hat{\theta}_9$ is

$$\begin{aligned}\mathrm{E}(\hat{\theta}_9) &= \mathrm{E}(e^{\bar{Y}})\mathrm{E}\left(e^{\frac{S^2(n-3)(n-1)}{2n(n-1)+2nS^2}}\right), \\ &= e^{\mu+\frac{\sigma^2}{2n}} \int_0^\infty e^{\frac{(n-1)(n-3)V}{2n(n-1)^2}+2nV} f(V)dV,\end{aligned}$$

where $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Therefore the bias and the MSE for $\hat{\theta}_9$ are

$$\mathrm{Bias}(\hat{\theta}_9) = \mathrm{E}(\hat{\theta}_9) - \theta \tag{17}$$

and

$$\mathrm{MSE}(\hat{\theta}_9) = \exp\left(2\mu+\sigma^2\right)\left(e^{[(2-n)/n]\sigma^2} f_1 - 2e^{[(1-n)/2n]\sigma^2} f_2 + 1\right), \tag{18}$$

where

$$\begin{aligned}f_1 &= \mathrm{E}\left[\exp\left(\frac{2(n-1)(n-3)S^2}{2n(n-1)+2nS^2}\right)\right], \\ f_2 &= \mathrm{E}\left[\exp\left(\frac{(n-1)(n-3)S^2}{2n(n-1)+2nS^2}\right)\right].\end{aligned}$$

The Bayes estimator under quadratic loss tends to underestimate the true value of the log-normal mean. It is very efficient for different sample sizes and returns a small MSE when σ^2 is large. This estimator is also recommended when MSE is used as the criterion.

3.2 Bayes Estimator Under Relative Quadratic Loss

In certain circumstance, we are interested in the Bayes estimator under quadratic loss. Enrico and Carlo [2] proposed another estimator under relative quadratic loss.

Define the parameter $\tau = -\ln(\theta)$. Therefore, $\theta^{-1} = \exp(\tau)$, $\theta^{-2} = \exp(2\tau)$. We have $\mathrm{E}(\theta^{-1}) = \mathrm{E}(e^{\tau})$ and $\mathrm{E}(\theta^{-2}) = \mathrm{E}(e^{2\tau})$. Since the distributions of τ and 2τ are known, using the same prior, the moment generating function can be used to find the expectation of θ^{-1} and θ^{-2}. Thus,

$$\tau|X \sim GH(\bar{\lambda}, \bar{\alpha}, \bar{\beta}, \bar{\delta}, \bar{\mu}),$$

and

$$2\tau|X \sim GH(\bar{\lambda}, \bar{\alpha}/2, \bar{\beta}/2, 2\bar{\delta}, -2\bar{\mu}),$$

where $\bar{\lambda}, \bar{\alpha}, \bar{\beta}, \bar{\delta}$ and $\bar{\mu}$ are defined as before.

The relative quadratic loss is $L = (\frac{\theta-\hat{\theta}}{\theta})^2 = (1 - \frac{\hat{\theta}}{\theta})^2$, and

$$\mathrm{E}(L|\sigma^2) = 1 - 2\hat{\theta}\mathrm{E}\left(\frac{1}{\theta}\right) + \hat{\theta}^2\mathrm{E}\left(\frac{1}{\theta}\right)^2.$$

Taking the derivative with respect to $\hat{\theta}$, we have

$$0 = 0 - 2\mathrm{E}\left(\frac{1}{\theta}\right) + 2\hat{\theta}\mathrm{E}\left(\frac{1}{\theta}\right)^2.$$

Hence the Bayes estimator under relative quadratic loss is

$$\begin{aligned}\hat{\theta} &= \frac{\mathrm{E}(\theta^{-1})}{\mathrm{E}(\theta^{-2})} = \frac{\mathrm{M}_{\mathrm{GH}}(\tau)}{\mathrm{M}_{\mathrm{GH}}(2\tau)} \\ &= \exp(a\bar{Y})\left(\frac{\frac{n}{a^2}(\gamma^2 - \frac{4a^2}{n} + 4b)}{\frac{n}{a^2}(\gamma^2 - \frac{a^2}{n} + 2b)}\right)^{(\lambda - \frac{n-1}{2})/2} \times \frac{K_{\lambda-\frac{n-1}{2}}\sqrt{(\gamma^2 - \frac{a^2}{n} + 2b)((n-1)S^2 + \delta^2)}}{K_{\lambda-\frac{n-1}{2}}\sqrt{(\gamma^2 - \frac{4a^2}{n} + 4b)((n-1)S^2 + \delta^2)}}.\end{aligned}$$

Note that if setting $b^\star = b - 2a^2/n$ and $\gamma^{\star 2} = \gamma^2 - 4a^2/n + 4b$, it is the same Bayes estimator obtained under quadratic loss. The approximate estimator is replacing b by $b^\star = b - 2a^2/n$ to the original estimator. So the estimator is

$$\hat{\theta}_{10} = \exp\left(\bar{Y} + \frac{S^2(n-7)(n-1)}{2n(n-1) + 2n\sigma^2}\right).$$

Replacing σ^2 by its unbiased estimator S^2, we have

$$\hat{\theta}_{10} = \exp\left(\bar{Y} + \frac{S^2(n-7)(n-1)}{2n(n-1) + 2nS^2}\right).$$

$$\mathrm{E}(\hat{\theta}_{10}) = \mathrm{E}\left(e^{\bar{Y}}\right)\mathrm{E}\left(e^{\frac{S^2(n-7)}{2n(n-1)+2nS^2}}\right),$$

$$= e^{\mu + \frac{\sigma^2}{2n}} \int_0^\infty e^{\frac{\frac{(n-1)(n-7)V}{2n(n-1)^2}}{\sigma^2} + 2nV} f(V)dV,$$

where $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

The bias and the MSE of this estimator are

$$\text{Bias}(\hat{\theta}_{10}) = \text{E}\left(\hat{\theta}_{10}\right) - \theta \tag{19}$$

and

$$\text{MSE}(\hat{\theta}_{10}) = \exp(2\mu + \sigma^2)\left(e^{[(2-n)/n]\sigma^2} f_1 - 2e^{[(1-n)/2n]\sigma^2} f_2 + 1\right), \tag{20}$$

where

$$f_1 = \text{E}\left[\exp\left(\frac{2(n-1)(n-7)S^2}{2n(n-1)+2nS^2}\right)\right],$$
$$f_2 = \text{E}\left[\exp\left(\frac{(n-1)(n-7)S^2}{2n(n-1)+2nS^2}\right)\right].$$

The Bayes estimator under relative quadratic loss tends to underestimate the ture value of the log-normal mean. It has a large MSE when the sample size is small. However, as the sample size gets large, it performs well regardless of the size of σ^2.

4 COMPARISON OF THE ESTIMATORS

In the previous sections, we summarized ten alternatives to estimate log-normal means, including frequentist methods and Bayesian methods. Each of the estimator has its advantages and disadvantages. Therefore, it is better to make a comparison of these ten estimators.

The relative bias using the true parameter θ as the reference, it allows one to check whether each estimator is overestimated or underestimated. The MSE criterion is used to compare estimators presented in Chapters. We will compare these methods using different values of σ^2 and sample size. We will look at values of σ^2 from 0.1 to 5 by increments of 0.1 and three different values for the sample size: 10, 50, and 100, which corresponds to a small moderate, and large sample size, respectively. The relative MSE is calculated using the "approximately minimum MSE estimator" ($\hat{\theta}_3$) as the reference. The reason to present the relative MSE is that we can eliminate the influence of the parameter μ, and thus we only need to consider the effect of σ^2.

Figures 2-4 present the relative MSE for different sample sizes. Figure 2 shows that for small sample size, most of the estimators have a MSE greater than or equal to the "approximately minimum MSE estimator" ($\hat{\theta}_3$). Two estimators, "a degree of freedom adjusted" MLE ($\hat{\theta}_8$), and "Bayes estimator under quadratic loss" ($\hat{\theta}_9$) have smaller MSE than the others.

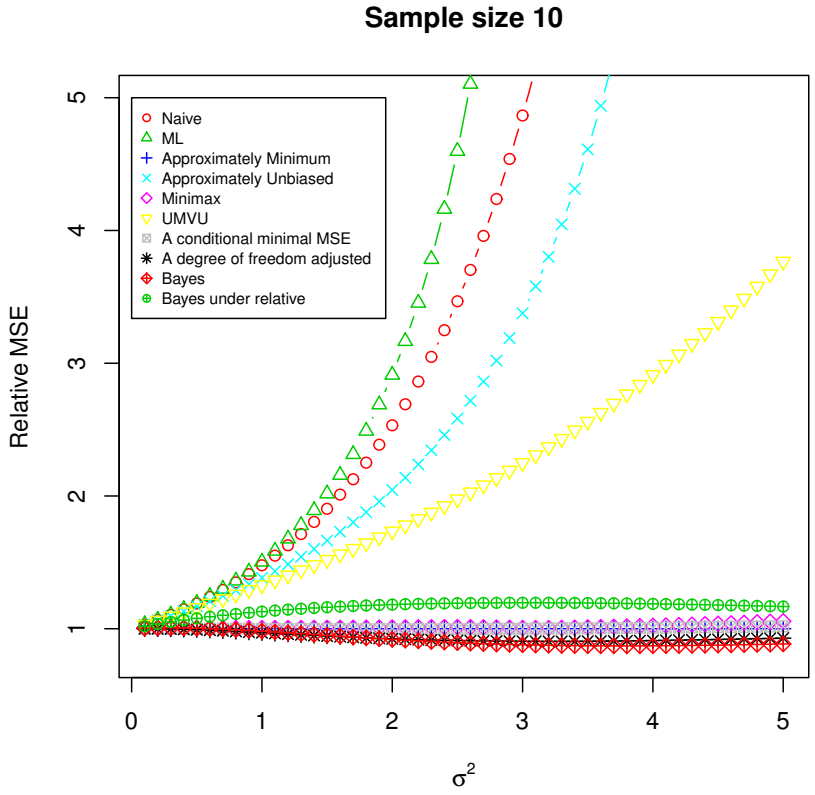

Figure 2: Relative MSE for sample size 10

As σ^2 increases, the "Bayes estimator under quadratic loss" shows some advantages. Although the "minimax estimator" ($\hat{\theta}_5$) and the "conditional minimal MSE estimator" ($\hat{\theta}_7$) are less inefficient than the previous two estimators, they had a smaller MSE. In addition, the "naive estimator" ($\hat{\theta}_1$) and "maximum likelihood estimator" ($\hat{\theta}_2$) are very inefficient compared to other estimators.

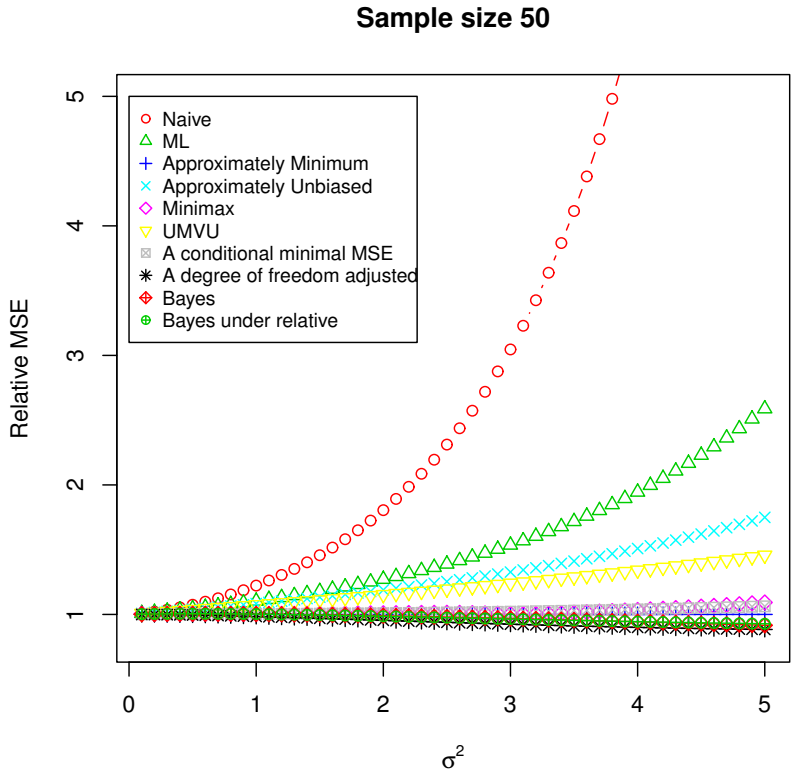

Figure 3: Relative MSE for sample size 50

Figure 3 shows that when the sample size increases to 50, the performance of the "naive estimator" ($\hat{\theta}_1$) is not influenced by the sample size; it still has a large MSE. The MSE of the"maximum likelihood estimator" ($\hat{\theta}_2$) begin to decrease. Other estimators start to come close to "approximately minimum MSE estimator" ($\hat{\theta}_3$).

Figure 4 indicates that when the sample size gets larger, the "naive estimator" ($\hat{\theta}_1$) still has a large relative MSE, while the other estimators become close to "approximately minimum MSE estimator" ($\hat{\theta}_3$). This indicates that when the sample size is very large, the difference among those ten estimators becomes smaller.

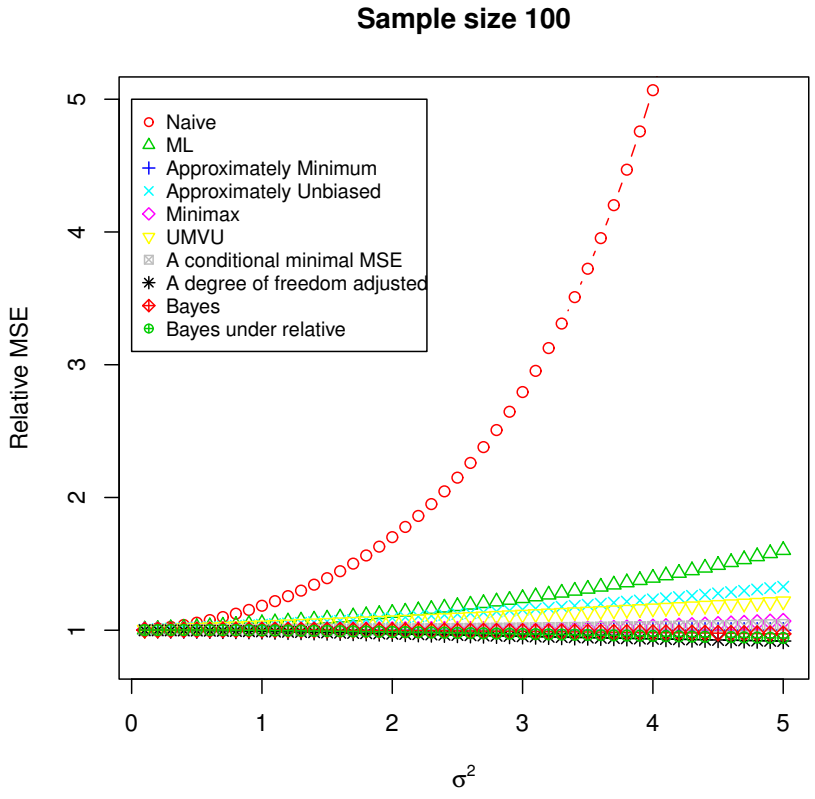

Figure 4: Relative MSE for sample size 100

Figures 5-7 display the relative bias for different sample sizes. For the relative bias of a small sample size, Figure 5 shows that only the "MLE" ($\hat{\theta}_2$) is over estimated and others are either unbiased or underestimated. The underestimated estimator are: "a conditional minimal MSE estimator" ($\hat{\theta}_7$), "a degree of freedom adjusted" MLE($\hat{\theta}_8$), "Bayes estimator under quadratic loss"($\hat{\theta}_9$) and "Bayes estimator under relative quadratic loss" ($\hat{\theta}_{10}$). The figures also show that the "naive estimator" ($\hat{\theta}_1$), "the approximately unbiased estimator" ($\hat{\theta}_3$) and "UMVUE" ($\hat{\theta}_6$) are the unbiased estimators, which are consistent with the theoretical results.

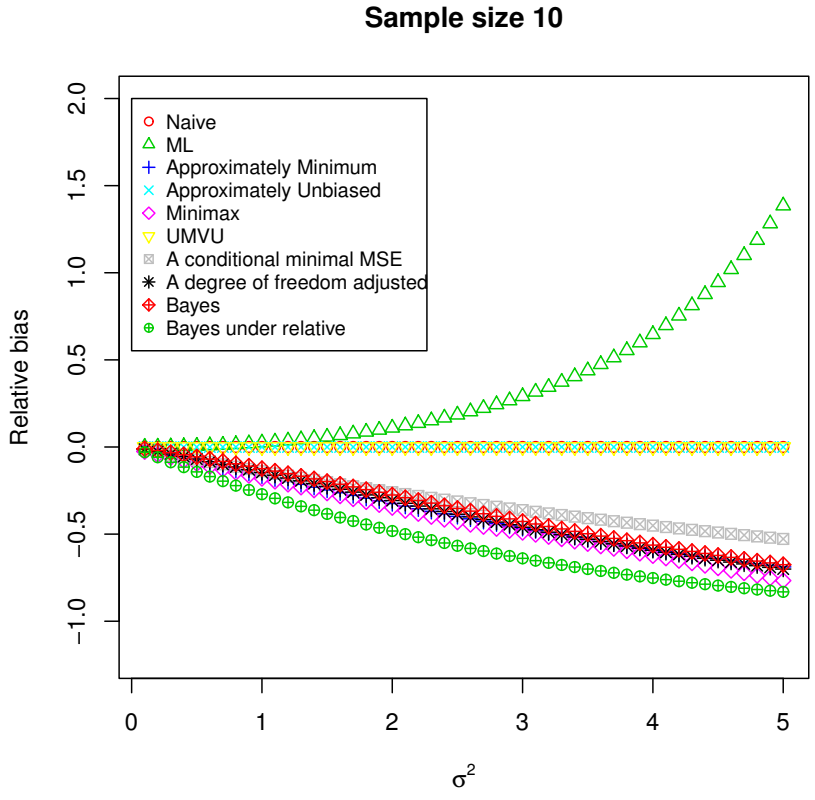

Figure 5: Relative bias for sample size 10

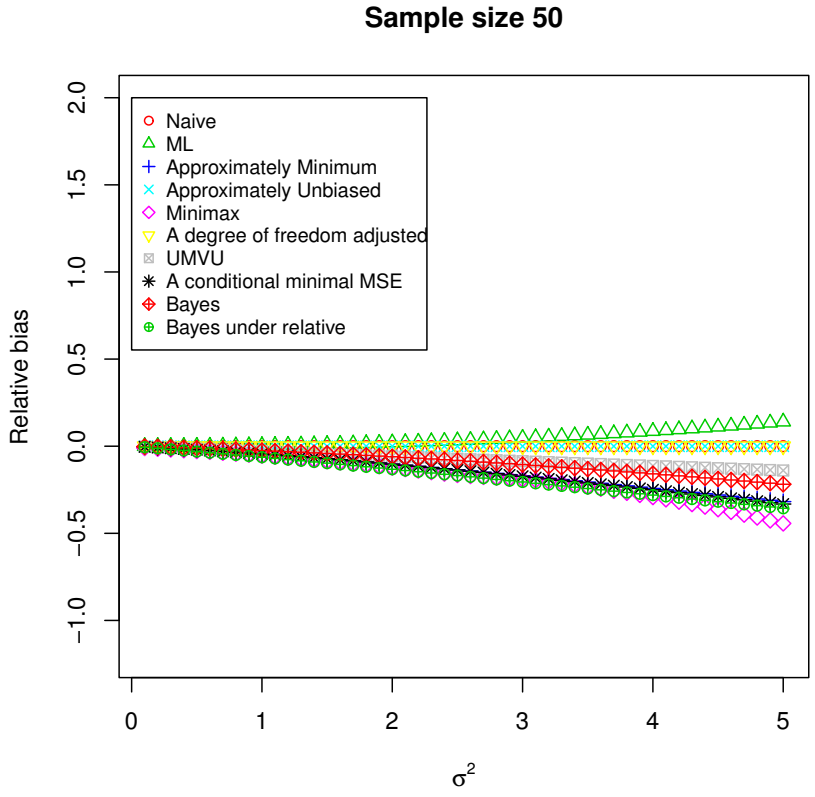

Figure 6: Relative bias for sample size 50

Figure 6 illustrates that when sample size increase to 50, the relative bias of the "maximum likelihood estimator" starts to decrease. The other underestimated estimators also begin approaching the true parameter.

We see from Figure 7 that as the sample size gets large, the relative bias for all the estimators are close to zero. This indicates that all of these estimators are less biased when the sample size is large. However, for the same sample size, all biased estimators have a large bias for large value of σ^2

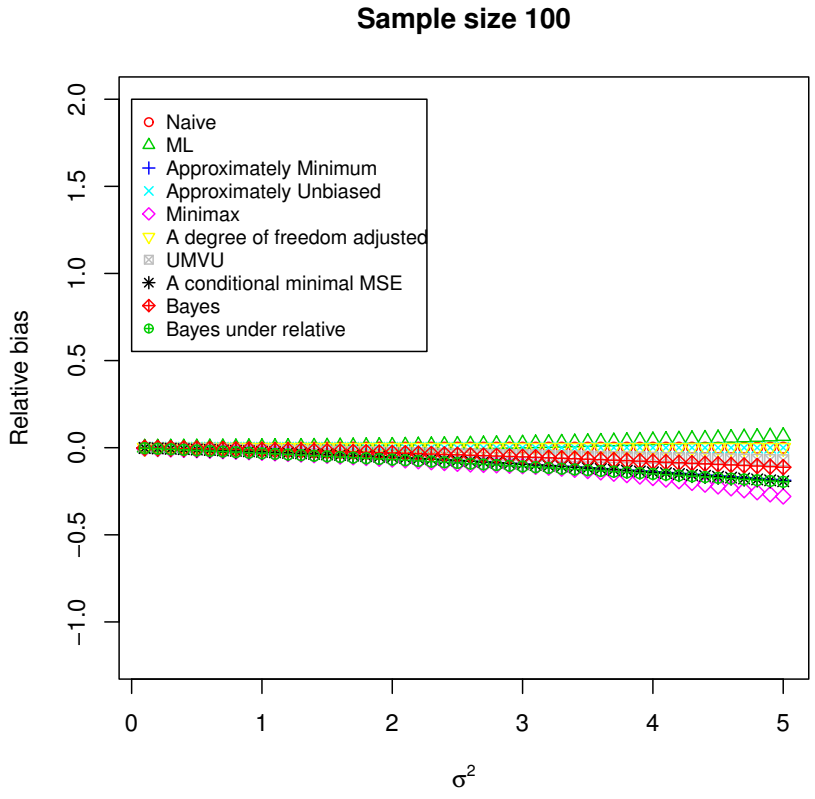

Figure 7: Relative bias for sample size 100

5 SIMULATION

Simulations were done to verify the theoretical results and to check any deviations when dealing with real data. In this simulation, we set μ=0 and σ^2 takes values from 0.1 to 5.0 with a segment of 0.1. The sample size was set to 10, 50 and 100, respectively. For each sample size and each σ^2, a random sample is drawn from the log-normal distribution and the ten estimates were calculated. The procedure was repeated 5000 times. The bias and the MSE of each estimator are calculated. Figures 7-12 portray the simulation results.

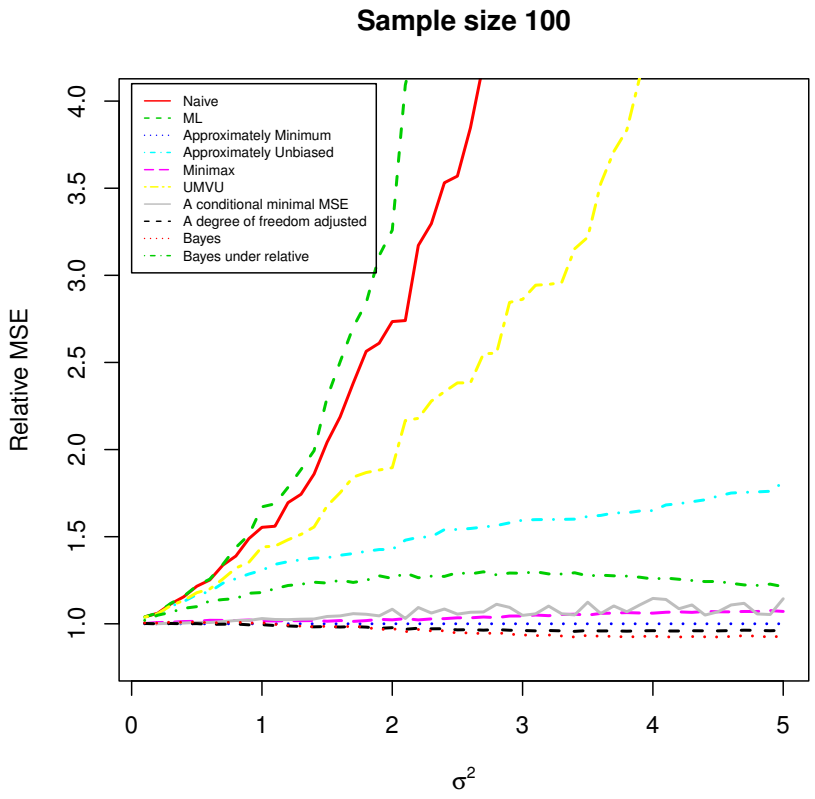

Figure 8: Simulations for relative MSE of sample size 10

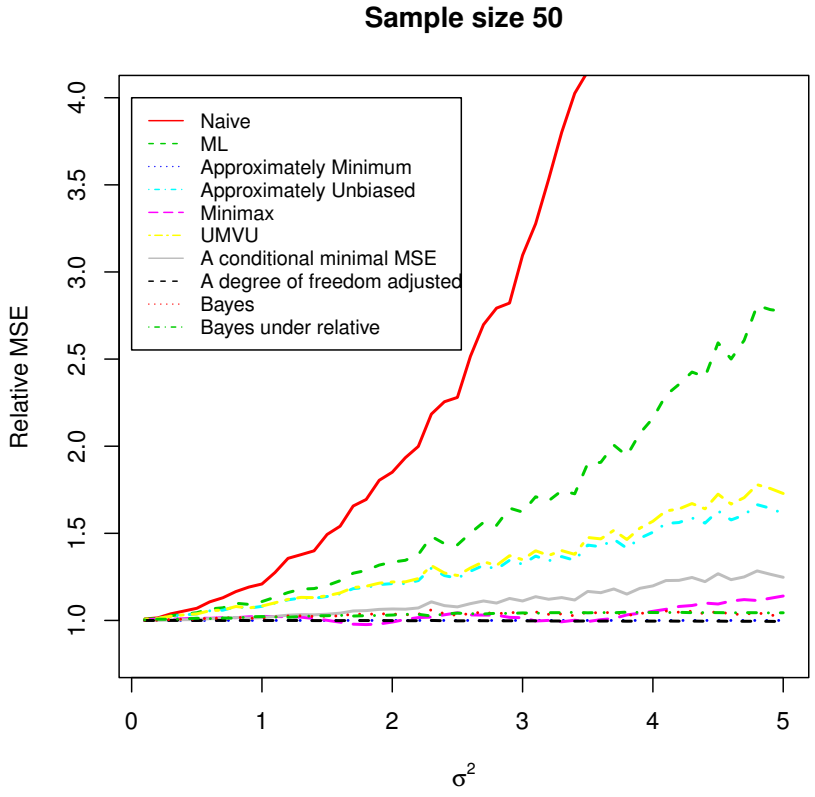

Figure 9: Simulations for relative MSE of sample size 50

Figures 8-10 depict the simulation results of relative MSE with different sample sizes. When the sample size is small, Figure 8 shows that the MSE of both "a degree of freedom adjusted" MLE ($\hat{\theta}_8$) and "Bayes estimator under quadratic loss" ($\hat{\theta}_9$) are smaller than the MSE of "approximately minimum MSE estimator" ($\hat{\theta}_3$). When σ^2 gets larger, "Bayes estimator under quadratic loss" ($\hat{\theta}_9$) has a smaller MSE than "a degree of freedom adjusted" MLE ($\hat{\theta}_8$). This is consistent with the theoretical results.

When the sample size increases to 50, Figure 9 indicates that the "naive estimator" ($\hat{\theta}_1$) still returns a large MSE.

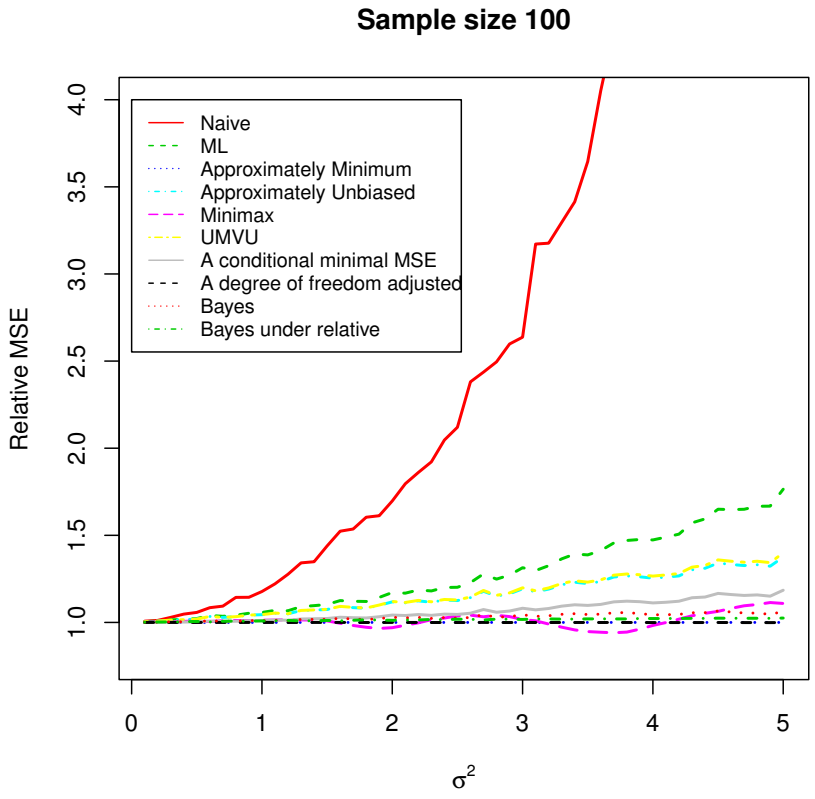

Figure 10: Simulations for relative MSE of sample size 100

As the sample size gets closer to 100, Figure 10 shows that most of the estimators MSE tend to approach the "approximately minimum MSE estimator" ($\hat{\theta}_3$) 's MSE except the "naive estimator" ($\hat{\theta}_1$). This verifies the theoretical results are correct when dealing with real data.

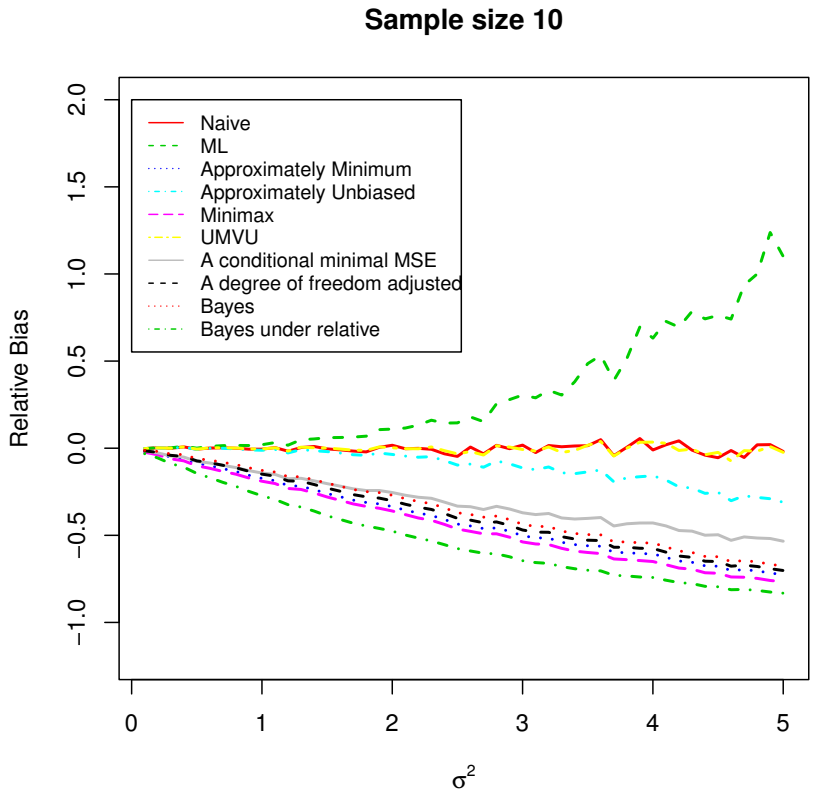

Figure 11: Simulations for relative bias of sample size 10

Figures 11-13 illustrate the simulation results of relative bias with different sample sizes. When the sample size is small, Figure 11 shows that the "MLE" ($\hat{\theta}_2$) is an over-estimated estimator, and "naive estimator" ($\hat{\theta}_1$), "approximately unbiased estimator" ($\hat{\theta}_3$) and "UMVUE" ($\hat{\theta}_6$) are all unbiased estimator. These results are consistent with the theoretical conclusions.

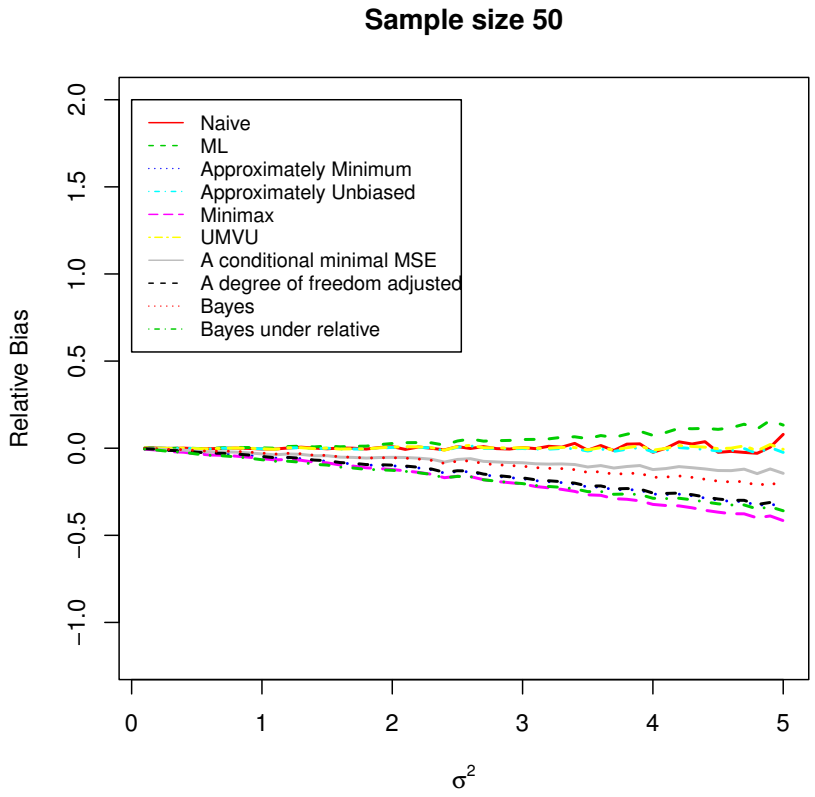

Figure 12: Simulations for relative bias of sample size 50

It can be seen from Figure 12 that when the sample size increases to 50, these ten estimators begin to approach the true parameter.

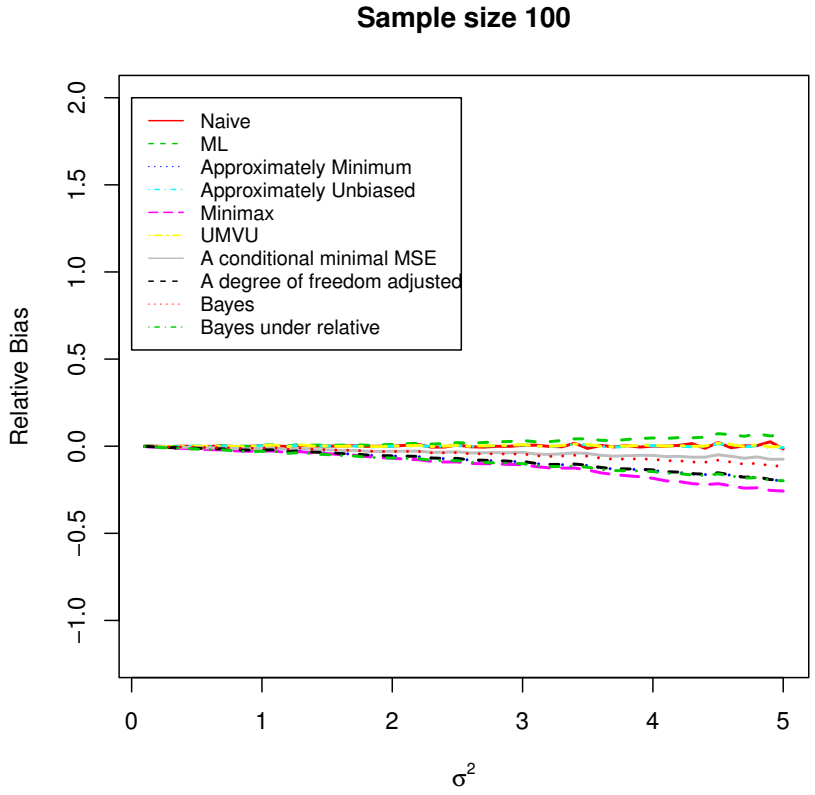

Figure 13: Simulations for relative bias of sample size 100

When the sample size is 100, Figure 13 illustrates that all of the estimators are approximately equal to the true mean value, which indicates that all of these estimators tend to be unbiased when sample size is large. This is also consistent with the theoretical results.

As we can see form the figures, the simulations curves are not as smooth as they are in the theoretical graphs. The reason is that the data are randomly generated from the log-normal distribution and the the results are based on limited number of repetitions. Thus, the curves in the simulations figures fluctuate around the theoretical lines.

6 APPLICATION

To present the application of the ten methods, we used the data from EPA [7], the Naval Construction Battalion Center (NCBC) Superfund Site in Rhode Island. The ground water samples were drawn from seventeen wells from the NCBC site. It was used for inorganic analyses. The purpose is to find the trustworthy estimates of the means of different inorganic contaminants at this area. For application, two contaminants, aluminum and manganese are analyzed.

The data for contaminants of aluminum are: 290, 113, 264, 2660, 586, 71, 527, 163, 107, 71, 5920, 979, 2640, 164, 3560, 13200, 125. The sample mean and the standard deviation for the original data are 1849.412 and 3351.273, respectively. The sample mean and standard deviation for the log-transformed data are 6.225681 and 1.659261 respectively.

The data for contaminants of manganese are: 15.8, 28.2, 90.6, 1490, 85.6, 281, 4300, 199, 838, 777, 824, 1010, 1350, 390, 150, 3250, 259. The sample mean and standard deviation for the original data are 902.2471 and 1189.489. For the log-transformed data, the sample mean is 5.912132 and the standard deviation is 1.567666 respectively.

Figure 14 shows the histograms of the two contaminants. One can see that the aluminum data has a longer tail than the manganese data. Thus it is more skewed than the manganese data. A Shapiro Wilks test will be used to test for normality of the data.

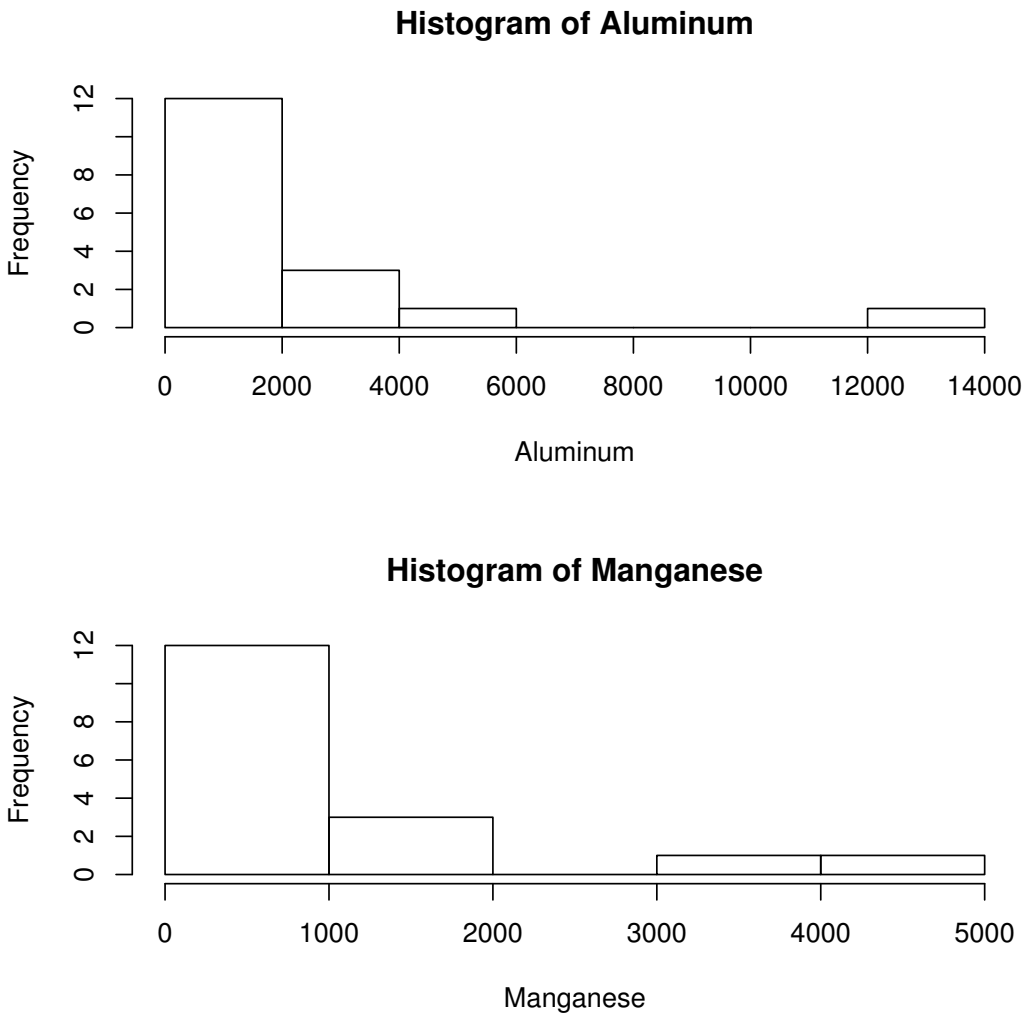

Figure 14: Histogram of two contaminants: aluminum and manganese

For the contaminants of aluminum, the Shapiro Wilks test yields a p-value of 0.000009173. This indicates that the original data is not normally distributed. For the natural logarithm of the data, the p-value is 0.1134. This means after the transformation, the data is normally distributed and the original data of the aluminum is log-normal distribution.

For the contaminants of manganese, the Shapiro Wilks test has a p-value of 0.000225 for the original data and p-value of 0.7994 for the natural logarithm of the data. Thus, the original data of the manganese has a log-normal distribution.

Ten methods were applied to these two datasets. The point estimates for the log-normal means of the two contaminants are presented in Table 1.

For aluminum, the order of the point estimates from the smallest to largest is: $\hat{\theta}_{10}$, $\hat{\theta}_5$, $\hat{\theta}_3$, $\hat{\theta}_8$, $\hat{\theta}_9$, $\hat{\theta}_7$, $\hat{\theta}_4$, $\hat{\theta}_6$, $\hat{\theta}_2$, $\hat{\theta}_1$. Both the "naive estimator" ($\hat{\theta}_1$) and the "maximum likelihood estimator" ($\hat{\theta}_2$) are large. This is because they are inefficient estimators, which tend to have a large estimates. Note that the "approximately minimum MSE estimator" ($\hat{\theta}_3$), "minimax estimator" ($\hat{\theta}_5$), "the conditional MSE estimator" ($\hat{\theta}_7$), "a degree of freedom adjusted" MLE ($\hat{\theta}_8$) and the "Bayes under quadratic loss" estimator ($\hat{\theta}_9$) have relatively small estimates. This observed results correspond with the theoretical conclusions.

For manganese, the order of the point estimates from the smallest to largest is: $\hat{\theta}_{10}$, $\hat{\theta}_5$, $\hat{\theta}_3$, $\hat{\theta}_8$, $\hat{\theta}_9$, $\hat{\theta}_1$, $\hat{\theta}_7$, $\hat{\theta}_4$, $\hat{\theta}_6$, $\hat{\theta}_2$. The results are similar to the contaminants of aluminum. However, the "naive estimator" ($\hat{\theta}_1$) is smaller than some of the others. The "maximum likelihood estimator" ($\hat{\theta}_2$) still returns a large value. "Bayes under relative quadratic loss" ($\hat{\theta}_9$) gives us small estimates in both cases.

Table 1: Point Estimates for the Log-normal means

	Aluminum	Manganese
Estimate		
$\hat{\theta}_1$	1849.41	902.2471
$\hat{\theta}_2$	1846.927	1174.548
$\hat{\theta}_3$	1178.845	797.2196
$\hat{\theta}_4$	1672.057	966.1599
$\hat{\theta}_5$	1112.763	747.2176
$\hat{\theta}_6$	1704.844	1100.925
$\hat{\theta}_7$	1372.127	905.091
$\hat{\theta}_8$	1214.563	819.38
$\hat{\theta}_9$	1329.953	888.3252
$\hat{\theta}_{10}$	1008.835	691.3925

7 CONCLUSION AND FUTURE WORK

In this thesis, we compared ten different estimating methods for log-normal means. For each method, the estimator, and its bias and MSE were given. Figures were produced based on the theoretical formula to compare the results visually. Simulations were done to support the theoretical result and to compare the results in the scenario of the real data.

As a result, "a degree of freedom adjusted" MLE ($\hat{\theta}_8$) and "Bayes estimator under quadratic loss" ($\hat{\theta}_9$) have a smaller MSE than the others. Although these two estimators are not unbiased estimator, they have some advantages. For large σ^2, the "bayes estimator under quadratic loss" ($\hat{\theta}_9$) is more efficient than "a degree of freedom adjusted" MLE ($\hat{\theta}_8$). To estimate log-normal means, "a degree of freedom adjusted" MLE ($\hat{\theta}_8$) is recommended when σ^2 is small and moderate, whereas "bayes estimator under quadratic loss" ($\hat{\theta}_9$) is favored when σ^2 is large.

There are several possible directions of future work. One possibility is to construct the confidence intervals for all of these estimators. The confidence interval is another criterion for measuring the accuracy of an estimator. It can also be used to compare the coverage of different methods. Bootstrapping method has been popular in calculating the confidence interval. It is a non-parametric method and easy to be applied to almost any problems and any datasets. Therefore, bootstrapping method may be added for further comparison.

In this thesis, only the estimators of log-normal means are discussed, so another possible direction is to find similar estimators of other log-normal measures, such as the median and mode, and then compare those against different sample sizes and σ^2.

BIBLIOGRAPHY

[1] Longford, Nicholas T. (2009). Inference with the Lognormal Distribution, *Journal of Statistical Planning and Inference*, 139, 2329-2340.

[2] Fabrizi, Enrico and Trivisano, Carlo (2012). Bayesian Estimation of Log-normal Means with Finite Quadratic Expected Loss, *International Society for Bayesian Analysis*, 7(5), 975-996.

[3] Zhou, XiaoHua (1998). Estimation of the Log-normal Mean, *Statistics in Medicine*, 17, 2251-2264.

[4] Shen, Haipeng, Brown, Lawrence D. and Zhi, Hui (2006). Efficient Estimation of Log-normal Means with Application to Pharmacokinetic Data, *Statistics in Medicine*, 25, 3023-3038.

[5] Rukhin, Andrew L. (1986). Improved Estimation in Log-normal Models, *Journal of the American Statistical Association*, 81(396), 1046-1049.

[6] Zellner, Arnold (1971). Bayesian and Non-Bayesian Analysis of the Log-normal Distribution and Log-normal Regression, *Journal of the American Statistical Association*, 66(334), 327-330.

[7] Singh, Ashok K., Singh, Anita and Engelhardt, Max (1997). The Lognormal Distribution in Environmental Applications, *Technology Suppport Center Issue*, 1-20.

[8] Office of Emergency and Remedial Response U.S. Environmental Protection Agency Washington, D.C. 20460 (2002). Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waster Sites, *OSWER*, 9285, 6-10.

[9] Finney DJ. (1941). On the Distribution of a Variate whose Logarithm is Normally Distributed. *Supplement to the Journal of the Royal Statistical Society*, 7, 144-161.

VITA

QI TANG

Education:	B.S. Statistics, North China of University Technology Beijing, China, July, 2011 M.S. Mathematics and Statistics, East Tennessee State University Johnson City, Tennessee, May, 2014
Professional Experience:	Graduate Assistant, East Tennessee State University, College of Arts and Science, 09/2012 - 05/2014