(DRAFT)

Quantification of Variability and Uncertainty Using Mixture Distributions: Evaluation of Sample Size, Mixing Weights and Separation between Components

Submitted to *Risk Analysis*

Junyu Zheng
H. Christopher Frey

Department of Civil Engineering
North Carolina State University
Raleigh, NC

ABSTRACT

Variability is the heterogeneity of values within a population. Uncertainty refers to lack of knowledge regarding the true value of a quantity. Mixture distributions have the potential to improve the goodness of fit to datasets not adequately described by a single parametric distribution. Uncertainty due to random sampling error in statistics of interests can be estimated based upon bootstrap simulation. In order to evaluate the robustness of using mixture distribution as basis for estimating both variability and uncertainty, 108 synthetic datasets generated from selected population mixture lognormal distributions were investigated, and properties of variability and uncertainty estimates were evaluated with respect to variation in sample size, mixing weight and separation between components of mixtures. Furthermore, mixture distributions were compared with single component distributions. Findings include: (1) mixing weight influences the stability of variability and uncertainty estimates; (2) bootstrap simulation results tend to be more stable for larger sample size; (3) when two components are well separated, the stability of bootstrap simulation is improved; however, a larger degree of uncertainty arises regarding the percentiles coinciding with the separated region; (4) when two components are not well separated, a single distribution may often be a better choice because it has fewer parameters and better numerical stability; and (5) dependencies exist in sampling distributions of parameters of mixtures and are influenced by the amount of separation between the components. An emission factor case study based upon NO_x emissions from coal-fired tangential boilers is used to illustrate the application of the approach.

KEY WORDS: Variability; uncertainty; mixture distributions; parameter estimation; bootstrap simulation

1. 0 INTRODUCTION

Variability is the heterogeneity of values with respect to different times, locations, or members of a population. Uncertainty, also referred to as fundamental or epistemic uncertainty, arises due to lack of knowledge regarding the true value of a quantity. [1, 2, 3, 4] Both variability and uncertainty may be quantified using probability distributions. Kaplan and Garrick [5] suggest that uncertainty regarding variability may be viewed in terms of probability regarding frequencies. Morgan and Henrion [6] and Frey [7] suggest that variability is described by a frequency distribution, and that uncertainty is described by probability distributions.

Probabilistic methods are being developed to quantitatively describe both variability and uncertainty. [8, 9, 10, 11, 12, 13] The recognition of the distinction between variability and uncertainty, especially with regard to potentially sensitive or highly exposed subpopulations in risk assessment, is growing. [14] There is a growing track record of the use of quantitative methods for characterizing variability and uncertainty in various applications including human health or ecological risk assessment and probabilistic emission estimation for various emission sources, including power plants, non-road mobile sources, and natural gas-fired engines.[11, 12, 15, 16, 17, 18]

A widely accepted method for uncertainty analysis is to identify inputs to a model that are known to have uncertainties, and to quantify the uncertainties in each input using a probability distribution model.[19, 20, 21] Commonly used probability distribution models include empirical and parametric distributions. In contrast to empirical distributions, parametric distributions allow for interpolation within the range of observed data and for extrapolation beyond the range of observed data to represent the tails of the distribution. The choice of empirical versus parametric distributions is not inherently a matter of right or wrong, but more a matter of preference of the analyst.[22] In practice, a parametric distribution is often used since it

is a compact means for representing variability in a quantity.

Single component distribution models such as the normal or lognormal distribution are often used to describe variability or uncertainty in a quantity. However, single component distributions might not well describe the variation in a quantity for some cases, such as when the data are actually based upon a mixture of two subpopulations. The use of single component distributions that are poor fits to data could lead to bias in variability and uncertainty analysis. A possible alternative is to use a finite mixture of distributions. A mixture distribution is comprised of two or more component distributions that are each weighted. Typically, a mixture distribution will produce a better fit to a data set than a single component distribution, because there are more parameters in the mixture distributions.

Mixture models have been used in the physical, chemical, biological and social science fields. For example, Harris [23] applied mixtures of geometric and negative binomial distributions to modeling crime and justice data. Kanji [24] described wind shear data using mixture normal distributions. Wedel *et al.*[25] utilized a finite mixture of Possison distributions to model the data on customer purchases of books offered through direct mail. In human exposure and risk assessment, Burmaster and Wilson[26] used mixture lognormal models to re-analyze data sets collected by the U.S. EPA for the concentration of $Radon^{222}$ in drinking water supplied from ground water, and found that the mixture model yielded an improved fit to the data not achievable with any single parameter distributions.

Frey and Rhodes [8, 9] presented a two-dimensional probabilistic approach for simultaneously quantifying variability and uncertainty based on single distributions featuring the use of bootstrap simulation. This general approach is adopted here. However, the approach is extended to include mixture distributions.

Because a mixture distribution has a more complicated mathematical form and more parameters than a single component distribution, the processes of parameter estimation and quantification of variability and uncertainty are more challenging. These challenges motivate the following key questions that are addressed by this paper:

1. How should the parameters of mixture distributions be estimated?
2. How should random numbers be generated from mixture distributions for purposes of bootstrap sampling?
3. How should confidence intervals be developed for statistics estimated from a mixture distribution?
4. How robust are results based upon mixture distributions with respect to sample size, mixing weight, and degree of separation between components of mixtures?
5. Under what circumstance is a single component distribution preferred over a mixture distribution?
6. What is the nature of the dependencies among the parameters of a mixture distribution?

This paper answers these six questions. In addition, the approach for using mixture distributions and bootstrap simulation is illustrated with a case study of an empirical dataset. This paper focuses on mixture lognormal distributions with two components since a lognormal distribution describes random variability resulting from multiplicative processes and often well describes the concentration of a chemical in the environment [26, 27]. However, the methods introduced here can be extended to other components and to mixture distributions with more than two components.

2. 0 METHODOLOGY

In this section, methods for fitting mixture distributions to data and methods for quantifying uncertainty in statistics estimated based upon mixture distributions are presented.

2.1 Definition of Mixture Distribution

According to the definition from Titterington *et al.* [28], a mixture model for a random variable or vector, X, is represented by a probability density function:

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x) \quad (1)$$

With

$$w_j > 0 \qquad \text{for } j=1,\ldots,k$$

And

$$w_1 + w_2 + \cdots + w_k = 1$$

Where,

f(x) = Probability density function for the mixture model

$f_k(x)$ = Probability density function (PDF) for a component of the mixture.

w_k = The mixing weight

k = number of components in the mixture

A mixture model with two components is expressed as:

$$f(x) = wf_1(x) + (1-w)f_2(x) \quad (2)$$

with $0 \le w \le 1$. For a two component lognormal distribution, $f_i(x)$ has the following form:

$$f_i(x) = \frac{1}{\sqrt{2\pi}\beta_i x} \exp\left[-\frac{(\ln(x) - \alpha_i)^2}{2\beta_i^2} \right] \quad (3)$$

Where,

α_i = The mean of *ln(x)* in the *i*th component of a mixture model

β_i = The standard deviation of *ln(x)* in the *i*th component of a mixture model

2.2 Parameter Estimation of Mixture Distributions

Many methods have been devised and used for estimating the parameters of a mixture distributions including, among others, Pearson's Method of Matching Moments (MoMM), informal graphical techniques and Maximum Likelihood estimation (MLE) approaches. [29] Until the use of computers became widespread in the 1960's, only fairly simple mixture density estimation problems were studied. MoMM has long been disfavored because of requirement that at least some useful statistics be known when estimating the parameters in a mixture models.[29, 30] However, this requirement cannot be met in many practical cases.

Prior to the widespread availability of computing resources, Cassie suggested graphical procedures employing probability paper as an alternative to moment estimates. [31] These graphical procedures work best on mixture populations that are well separated in the sense that each component has an associated region in which the presence of the other components can be ignored. [29, 30]

With the advent of high-speed computers, interest turned to likelihood estimation of the parameters in a mixture distribution. The general idea behind MLE is to choose values of the parameters of the fitted mixture distribution so that the likelihood that the observed data is a sample from the fitted distribution is maximized. [32]

MLE is selected as the preferred method for estimating parameters in a mixture distribution due to its relative practicality and generality. In MLE, the likelihood function is calculated by evaluating the probability density function for each observed data point conditioned on assumed parameter values and multiplying the results. [32] Alternatively, and

more commonly, the log-transformed version of the likelihood function is used. The MLE parameter estimates can be obtained by finding the maximum of a log-likelihood function through the use of a numerical analysis approach since analytical solutions are often not available for mixture distributions.

The log-likelihood function of a univariate mixture distribution is given by:

$$L=\sum_{i=1}^{n}\ln\left[f\left(x_i\middle|w,\alpha,\beta\right)\right]=\sum_{i=1}^{n}\ln\left[\sum_{j=1}^{k}w_j f_j\left(x_i\middle|\alpha_j\beta_j\right)\right] \quad (4)$$

Where,

$$\sum_{j=1}^{k}w_j=1$$

n = the number of data points

k = the number of components in a mixture distribution

L, = Log-likelihood function

α_j,β_j = the parameters in the j^{th} component in a mixture distribution

There are alternative approaches that can be used to find the maximum of Equation (4) and, hence, obtain the parameter estimates of a mixture distribution. One is the Expectation-Maximization (EM) algorithm.[33] The EM algorithm has the advantage of reliable global convergence, low cost per iteration, economy of storage and ease of programming; however, its convergence can be very slow in simple problems that are often encountered in practice,[29] and its results are strongly dependent upon the initial guesses assumed for the parameters.[30] A second approach is the Newton-Raphson iterative scheme. This scheme requires calculation of the inversion of the matrix of second derivatives of the log-likelihood function, which is complicated and must be done separately for each combination of parametric distributions assumed in a mixture (e.g., normal, lognormal, gamma, Weibull) thereby limiting general

applicability.[29, 30] A third approach preferred here because of its computational efficiency is to use nonlinear optimization methods to directly maximize the log-likelihood function.

The optimization problem for a mixture of two lognormal distributions is:

$$\begin{aligned} &\text{Maximize} && L = \sum_{i=1}^{n} \ln\left[w\, f_1(x_i|\alpha_1,\beta_1) + (1-w) f_2(x_i|\alpha_2,\beta_2)\right] \\ &\text{Subject to} && 0 \le w \le 1 \\ & && \alpha_1, \beta_1 > 0 \\ & && \alpha_2, \beta_2 > 0 \end{aligned} \tag{5}$$

This optimization problem is multidimensional and constrained. A variety of methods are available to solve such problems. Although an algorithm which can deal with a constrained problem has theoretical appeal, in practice, unconstrained methods are often easier to implement and provide robust results. Optimal results are checked against the constrained conditions. Those results that cannot meet the constraints can be abandoned and replaced during bootstrap simulation; however, this phenomenon occurs rarely in most cases as described later. Common unconstrained methods include the downhill simplex method; the direction-set method, of which Powell's method is the prototype; and others.[34] Powell's method is employed because it can provide reasonable estimation results and it is relatively easy to implement.

2.3 Quantification of Uncertainty in Statistics of Interests Using Mixture Distribution

Uncertainty in a statistic attributable to random sampling error can be represented by a sampling distribution.[20] Sampling distributions are used to estimate confidence intervals for the parameters of a distribution. A confidence interval for a statistic is a measure of the lack of knowledge regarding the value of the statistic. There are a variety of methods for characterizing uncertainty in statistics such as the mean or standard deviation, including analytical solutions and numerical simulations. Analytical solutions are available for cases in which the underlying

distribution for a data set is normal or for which the variance is small enough and/or the sample size for a data set is large enough (e.g., >30). If the underlying population distribution is not normal and the sample size for a data set is small, analytical methods based upon normality may lead to significant errors in the estimation of confidence intervals. Therefore, there is a need for a more flexible approach for estimating sampling distributions and confidence intervals when mixture distributions are used.

Bootstrap simulation, introduced by Efron in 1979, is a numerical technique originally developed for the purpose of estimating confidence intervals.[35] This method can provide solutions in situations where exact analytical solutions may be unavailable and in which approximate analytical solutions are inadequate.[20] Bootstrap simulation has been widely used in the prediction of confidence intervals for a variety of statistics.[8, 9, 20, 36, 37, 38]

In using bootstrap simulation, there are two major aspects. The first is a procedure for generating random samples from an assumed population distribution, and the second is the method of forming confidence intervals for statistics estimated from the random samples.[35, 36]

While there are standard numerical methods for drawing random samples from single component parametric distributions,[6, 20] the methods for drawing random samples from mixture distributions are more complicated in the context of bootstrap simulation. Although it is possible to obtain a single random sample from a mixture distribution by sampling from a weighted proportion of single component distributions, one of the objectives in bootstrap simulation is to develop confidence intervals for all statistics, including the component weights. Therefore, it is necessary to develop an estimate of the assumed population distribution in a manner that allows for the weight to vary randomly from one bootstrap sample to the next. For this purpose, an empirical distribution is used to represent the assumed population distribution for the mixture.

As shown in Figure 1, the first step in developing the assumed population distribution is to generate a large number of random samples using standard simulation methods. For example, suppose there is a mixture of two lognormal components, one with a weight of 40 percent and the other with a weight of 60 percent. In order to develop a stable and precise estimate of the cumulative distribution function (CDF) of this mixture, one may simulate 2,000 or more random values. Thus, on average 800 values would be simulated from the first component and 1,200 values would be simulated from the second component. These values would be rank-ordered to describe an empirical cumulative distribution function of the mixture distribution.

Once an empirical representation of the assumed population mixture distribution is available, it is then possible to randomly sample from it to generate bootstrap samples, as indicated in Figure 1. From each bootstrap sample, the bootstrap replicates of the component parameter values and of the weight may be estimated. For each bootstrap replication of the distribution parameters, the mean and other statistics may be simulated.

There are several methods for forming bootstrap confidence intervals such as the percentile, hybrid, bootstrap-t, and Efron's BC_a methods.[35, 37] The percentile method is perhaps the most frequently used in practice.[35, 38] The intervals from this method are the simplest to obtain, use and explain. The Hybrid method is justified by asymptotic results for the bootstrap in complicated models.[35, 38] The bootstrap-t and the BC_a intervals are comparable in that both have been demonstrated theoretically to be "second-order correct" for one-sided intervals in some relatively simple situations.[35, 38] The BC_a method was recommended for general use, especially for nonparametric problems;[35] however, the process for estimating BC_a confidence intervals is complicated and the computation burden is heavy.[38] Thus it appears that it is seldom used in actual applications. Therefore, for simplicity and because it is the most widely used

method in practice, the percentile method is used here to construct bootstrap confidence intervals.

3.0 PARAMETRIC STUDY OF VARIABILITY AND UNCERTAINTY BASED UPON MIXTURE DISTRIBUTIONS

In order to answer the fourth motivating question, synthetic datasets with different sample sizes, mixing weights and magnitudes of separation between components were generated from mixture lognormal distributions with two components. The assumed population mixture lognormal distributions are described in Table 1.

Twelve groups of population mixture distributions were evaluated. The groups differ in terms of the variability of each component and the relative degree of separation of the means of the two components compared to the variability. For example, when μ_1=1.0, σ_1=0.5, μ_2=1.5, σ_2=0.5, the means of the two components are separated by only one standard deviation and therefore are said not to be well separated. However, when μ_1=1.0, σ_1=0.5, μ_2=6.0, σ_2=0.5, the means of the two components are separated by 10 standard deviations and are said to be well separated.

To investigate the effect of variation of mixing weights and sample sizes, first component weights of 0.1, 0.3 and 0.5, and sample sizes of 25, 50 and 100 were studied parametrically for each group. Therefore, there are 9 synthetic datasets for each of the 12 groups, for a total of 108 synthetic datasets.

3.1 Evaluation of Robustness of Bootstrap Simulation Results

To illustrate the results obtained from the parametric study of different variability in components, separation between components, mixing weights and sample sizes, a series of results based upon Groups 5, 6, 7 and 8 are displayed graphically in Figures 2, 3, 4 and 5 for 1, 2, 4 and 10 standard deviation separation between the means of the two components, respectively.

Each component in all four figures has a standard deviation of 0.5. The variation in sample size and weight is illustrated by the panels of each figure. Each panel displays the assumed population distribution and the 95 percent confidence interval on the CDF. The later was obtained based upon bootstrap simulation.

Figure 2, which represents a situation with relatively little separation between components, illustrates that the MLE optimization method works better for larger sample size and for large weights for both components. Conversely, the method failed to provide parameter estimates for small sample sizes and/or small weights for the first component. In situation where parameter estimation fails for a bootstrap sample, the bootstrap sample is abandoned and replaced with a new one. If this problem occurs ten times, the bootstrap simulation is judged to fail and no results are reported. For the successful cases it is clear that the 95 percent confidence intervals become narrower for a given component weight as the sample size increases. It is difficult to visually detect much difference in the population distribution or the confidence intervals when comparing different weights for the same sample size.

Figure 3, 4 and 5 illustrate that the parameter estimation method becomes more robust as the degree of separation increases, but that there are still combinations of small sample size (i.e., 25 and 50) for the smallest weight considered for which the bootstrap results fail. However, bootstrap results were obtained for cases of n=25 and w=0.3, and n=100 and w=0.1, which failed for the smallest separation among components as shown in Figure 2. In general, the width of the confidence intervals decreases as sample size increases. Although some separation of components is evident in Figure 3, the separation is highly pronounced in Figures 4 and 5, as revealed by the well-defined inflection points. The inflection points occur at a cumulative probability equal to the corresponding weight. For example, as shown in Figure 5, the inflection

point for a mixture distribution with a weight of 0.5 occurs near the median of the CDF.

As the degree of separation increases, the range of uncertainty for percentiles of the distribution near the inflection point increases. Because the weight parameter is itself a random variable, there is uncertainty regarding the inflection point, leading to widening of the confidence intervals around this point. Thus, even in Figure 3 where the separation is relatively small, there is a noticeable "bulge" in the confidence intervals near the inflection point, especially for the small weights.

The results not graphically shown in this paper for case studies with $\sigma_1=\sigma_2=0.1$ and $\sigma_1=\sigma_2=1.0$, yielded similar characteristics to the results with $\sigma_1=\sigma_2=0.5$ shown here.

4.0 COMPARISONS BETWEEN SINGLE DISTRIBUTIONS AND MIXTURE DISTRIBUTIONS

For comparison purposes, single lognormal distributions were fit to the datasets generated from the specified mixture population distributions listed in Table 1. Selected results comparing the 95 confidence intervals of the CDF of a single lognormal distribution to the population mixture distribution are shown in Figure 6 and 7 for cases of one and two standard deviation separation between components of the population mixture, respectively.

Figure 6 illustrates that a single component distribution may appear to provide a good fit to sample data from a population mixture distribution particularly if each component has a large weight, if the components are not well separated, and if the sample size is small. For example, for the case of n=25 and w=0.5, the confidence interval for the single component distribution well encloses the population mixture distribution. As the sample size increases, bias near the lower tail and the median become more pronounced. For the small weights, biases are evident in the lower tail, which corresponds to the component with less weight. As the degree of separation

increases, biases associated with fitting a single component distribution become more pronounced, as shown in Figure 7. In all cases, the tails and the inflection region are poorly represented. With an increase in the separation of the components, the use of single component distributions became clearly unreasonable.

Quantitative summaries of results for both fitted mixture and fitted single component distributions are given in Tables 2 and 3 for w=0.3 and w=0.5, respectively. Both tables are based upon n=100. The point values of the 2.5, 30, 50, 75 and 97.5 percentiles, and of the mean and standard deviation, of the population distribution are given in each case. The 95 percent confidence intervals for these statistics are also given based upon the results of bootstrap simulation.

In the case of the fitted mixture distributions, the confidence intervals enclose the population values in all cases. However, when a single component distribution is used, biases are more pronounced for the smaller weight, increased separation between components, and increased variability for each component, and are particularly evident for percentiles near the inflection point of the mixture distributions. Conversely, the results suggest that it may be difficult to discern that a single component is not appropriate if the components are not well separated and/or if the variability within the components is relatively small.

5.0 DEPENDENCIES AMONG SAMPLING DISTRIBUTIONS OF PARAMETERS OF MIXTURE DISTRIBUTIONS

The dependencies among estimated parameters of a fitted mixture distribution were investigated. Figures 8, 9, and 10 display scatter plots of bootstrap simulation results for parameters of two component lognormal distributions with separation between components of two, four and ten standard deviations, respectively, based upon n=100 and w=0.5.

From Figure 8, there exists a positive dependency between parameters in most cases except for β_2, for which there is negative dependency versus w, α_1, α_2, β_1. For example, with an increase in w, there are more samples with values that fall within the first component and fewer samples that fall within the second component. Thus, there will be an increase in the means of both components and in the standard deviation of the first component, simultaneous with a decrease in the standard deviation of the second component because variability in the second component is reduced.

When there is large separation between the two components, the variation in the mixing weight will not lead to a substantial increase or decrease of mean in a component; hence, the dependency between the parameters become weaker. For example, the correlation coefficients between w and α_1, w and β_1, w and α_2, w and β_2 for the case of Figure 10 are -0.006, -0.02, 0.005 and 0.04, respectively. These results indicate that the dependencies among parameters in a mixture distribution are strongly associated with the magnitude of separation between two components.

6.0 AN ILLUSTRATIVE CASE STUDY: NO_X EMISSION FACTOR FOR A COAL-FIRED POWER PLANT

The methodology for simulating variability and uncertainty based upon mixture distributions is demonstrated via an empirical case study of a 12-month average emission factor for a tangential-fired, coal-fired boiler with low NO_x burners and overfire air.[39] This dataset cannot be fit well by any single distributions and the number of data points in the data set is relatively small (n=36).

A mixture distribution with two lognormal components was fit to the case study dataset. The parameter estimation results for the mixture of two lognormal distributions are:

Mixing weight=0.352
1st component: Mean of ln(x)=6.064, Standard deviation of ln(x)=0.345
2nd component: Mean of ln(x)=6.269, Standard deviation of ln(x)=0.0847

The fitted mixture distribution is shown in Figure 11 in comparison to the data, as are the results of bootstrap simulation. The fitted lognormal distribution and the results of bootstrap simulation are shown in Figure 12 as a comparison example. In Figure 11, all of the data are within the 95 percent confidence interval, and approximately 92 percent of the data are within the 50 percent confidence interval indicating an excellent fit. In particular, there is good agreement between the right tail of the mixture distribution and the observed data. In Figure 12, approximately 84 percent of the data are within of the 95 percent confidence interval and 33 percent of the data are within the 50 percent confidence interval, indicative of a poor fit.

It is typically the case that the confidence interval for a positively skewed fitted single component distribution is widest at the upper percentiles of the distribution. However, in the case of the fitted mixture distribution, there is also a widening of the confidence interval at a cumulative probability between approximately 0.07 and 0.45. Table 4 shows estimates of uncertainty in the parameters of the fitted mixture distribution. The 95 confidence interval of the weight parameter is from 0.078 to 0.523. The range of uncertainty in the weight parameter causes the 'bulge' in the confidence interval of the fitted mixture distribution for the cumulative probabilities similar to the range of uncertainty in the weight.

Tables 5 and 6 summarize the results for the 95 percent confidence intervals of the mean and 95th percentile of variability, respectively, based upon the four distributions fit to the dataset. The results are based upon the average of 10 bootstrap simulations, each with 500 bootstrap samples. The numerical precision of the estimates is indicated by intervals given in brackets. These intervals were estimated based upon the standard error of the ten bootstrap simulations.

The standard error is typically less than 0.5 percent of the mean value, indicating that the results are precise to almost three significant figures.

The upper bound of the 95 percent confidence interval for the mean based upon the mixture distribution, which has a best estimate of 532 and a precision of 530 to 535, is significantly lower than the values based upon the single component distributions. For example, the corresponding estimate based upon the Weibull distribution is 543 with a precision of 542 to 545. Although the precision intervals for the lower bound estimates of the 95 percent confidence intervals overlap, it appears that the mixture distribution implies a higher value of this quantity than do the single component distributions. Therefore, the 95% confidence interval for the mean is significantly narrower when estimated based upon the mixture distribution compared to the single component distributions. The mixture distribution has a better fit to the upper tail of the empirical distribution of the data.

The differences between the mixture distribution and the single component distributions are more pronounced with respect to the 95 percent confidence interval for the 95^{th} percentile of variability. The mean estimate of this statistic is 638, with a precision of 631 to 645, based upon the mixture distribution. In contrast, the mean estimate is significantly higher when based upon any of the three single component distributions. The lower bound of the confidence interval based upon the mixture distribution is 581 with a precision of 578 to 584. This is substantially lower than for any of the single component distributions. The upper bound of the confidence interval is 750 with a precision of 739 to 762. This is significantly lower than for the other distributions. Thus, the mean and the confidence interval of the mean for the 95^{th} percentile have significantly lower values for the mixture distribution compared to any of the single component distributions. This is because the mixture distribution better fits the data in the upper tail and

does not overestimate the tail.

7.0 DISCUSSION

The use of mixture distributions is a promising means to improve the estimates of uncertainty in statistics estimated from the fitted distribution because of improved fit to data, compared to single component distributions. However, there is a "bulge" in the confidence interval in the region of cumulative probability representing the inflection point between components of the mixture. Thus, there is a clear trade-off between improved fit based upon an increased number of parameters and the range of uncertainty for at least portions of the CDF. As Leoroux [40] points out, the elimination of unnecessary components in a mixture might lead to more precise estimates of the parameters, and, by extension, of other statistics. Thus, it is important to have as many components in the mixture as needed to obtain a reasonable fit to the data, but not to have too many.

Aside from the "bulge", the confidence interval can be relatively narrow for other portions of the cumulative distribution and for some statistics. In the empirical case study, the narrowest confidence interval for the mean was obtained from the mixture distribution. Thus, a mixture distribution may yield the most statistically efficient estimate of the sampling distribution of the mean. The results would vary in other cases. For example, if the weight parameter led to an inflection point at a location similar to the mean, the confidence interval for the mean could be comparatively wide.

8.0 CONCLUSIONS

Mixture distributions have the potential to improve the goodness of fit to datasets not adequately described by a single parametric distribution. This paper successfully demonstrated methods for fitting mixture distributions to data and for making inferences regarding uncertainty

using illustrative two-component mixture lognormal distributions. The methods introduced here can be extended to other components, more components, or both.

MLE method is preferred for parameter estimation of mixture distributions because of its practicality and generality. For purpose of bootstrap simulation, a high-resolution empirical distribution is recommended to represent the assumed population distribution for the mixture distribution because it allows for the weight to vary randomly from one bootstrap sample to the next. Bootstrap simulation is recommended for use in developing confidence intervals for statistics estimated from mixture distributions since there are no analytical solutions available.

The robustness of variability and uncertainty analysis based upon mixture distribution with respect to sample size, mixing weight, and degree of separation between components of mixtures was evaluated. Results are more robust when components are of comparable weight or the sample size is sufficiently large. When two components are well separated, the stability of results is improved; however, larger uncertainty arises around the separated region.

A single component distribution may appear to provide a good fit to sample data from a population mixture distribution particularly if each component has a large weight, if the components are not well separated, and if the sample size is small. Under these circumstances, a single component distribution may be a better choice because it has fewer parameters and better numerical stability.

Substantial dependencies exist in the sampling distributions of the five parameters in cases with little separation between the components. However, with an increase of the magnitude in the separation, the parameters become independent.

Recommended future studies include the extension of the approaches presented here to other types of components or to mixture distributions with more components. The percentile

method was used here to form bootstrap confidence intervals for both single component and mixture distributions. There may be opportunities to obtain improved results with other methods, such as the BC_a method, with a trade of increased computational complexity.

The use of mixture distributions is a promising method for improving the fit of distributions to data and for obtaining improved estimates of uncertainty in statistics estimated from the fitted distribution. The use of mixture distributions should be considered and evaluated in situations in which single component distributions are unable to provide acceptable fits to the data, or in situations in which it is known that the data arise from a mixture of distributions.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) grants program via grants R826766 and R826790. This paper has not been subject to any EPA review. Therefore, it does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

REFERENCE

1. Bogen, K.T. and Spear, R.C., "Integrating uncertainty and interindividual variability in environmental risk assessment," *Risk Analysis* **7**, 427-436 (1987).

2. IAEA (International Atomic Energy Agency), "Evaluating the reliability of predictions made using environmental transfer models," *Safety Series*, No.100, Vienna, Australia, International Atomic Energy Agency (1989).

3. Hattis, D. and Burmaster, D.E., "Assessment of variability and uncertainty distributions for piratical risk analyses," *Risk Analysis* **14**, 713-729 (1994)

4. Haimes, Y.Y, Barry,T., and Lambert, J.H., Eds., "Workshop proceedings: When and how can you specify a probability distribution when you don't know too much?" *Risk Analysis* **14**, 661-706 (1994).

5. Kaplan, S. and Garrick, B.J., "On the quantitative definition of risk," *Risk Analysis* **1**, 11-27 (1981).

6. Morgan, M.G., and M. Henrion, *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis* (Cambridge University Press, New York, 1990).

7. Frey, H.C., 1992, "Quantitative Analysis of Uncertainty and Variability in Environmental Policy Making," Directorate for Science and Policy Programs, American Association for the Advancement of Science, Washington, DC.

8. Frey, H.C. and Rhodes, D.S., "Characterization and Simulation of Uncertain Frequency Distributions: Effects of Distribution Choice, Variability, Uncertainty, and Parameter Dependence," *Human and Ecological Risk Assessment* **4**, 423-468 (1998).

9. Frey, H.C. and Rhodes, D.S.," Characterizing, Simulating and Analyzing Variability and Uncertainty: An illustration of Methods Using an Air Toxics Emissions Example," *Human and Ecological Risk Assessment* **2**, 762-797 (1996).

10. Boyce, C.P., "Comparison of Approaches for Developing Distributions for Carcinogenic Slope Factors," *Human and Ecological Risk Assessment* **4**, 527-577 (1998).

11. Kelly, E.J., Roy-Harrison, W., "A Mathematical Construct for Ecological Risk: A Useful Framework for Assessments," *Human and Ecological Risk Assessment* **4**, 229-244 (1998).

12. Cohen, J.T., Lampson, M.A., and Bowers, S., "The Use of Two-stage Monte Carlo Simulation Techniques to Characterize Variability and Uncertainty in Risk Analysis," *Human and Ecological Risk Assessment* **2**, 939-971 (1996).

13. Goodman, D., "Extrapolation in Risk Assessment: Improving the Quantification of Uncertainty, and Improving Information to Reduce the Uncertainty," *Human and Ecological Risk Assessment* **8**, 177-192 (2002).

14. Helton, J.C., J.D. Johnson, Jow, H.-N, McCurley, R.D., and Rahal, L.J., "Stochastic and Subjective Uncertainty in the Assessment of Radiation Exposure at the Waste Isolation Pilot Plant," *Human and Ecological Risk Assessment* **4**, 469-526 (1998).

15. Frey, H.C., R. Bharvirkar, J. Zheng, "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation," Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, pp 1-10, (1999).

16. Frey, H.C., and Bammi, S., "Quantification of Variability and Uncertainty in Lawn and Garden Equipment NO_x and Total Hydrocarbon Emission Factors," *Journal of the Air & Waste Management Association*, **52**, 435-448 (2002).

17. Frey, H.C., and S. Li, "Quantification of Variability and Uncertainty in Natural Gas-fueled Internal Combustion Engine NO_x and Total Organic Compounds Emission Factors," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001).

18. Zheng, J., H.C. Frey, "Quantitative Analysis of Variability and Uncertainty in Emission Estimation: An Illustration of Methods Using Mixture Distributions," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001).

19. Frey, H.C., Zheng, J., "Probabilistic Analysis of Driving Cycle-Based Highway Vehicle Emission Factors," (Accepted for publication, August, 2002, *Environmental Science and Technology*).

20. Cullen, A.C., and Frey, H.C., *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, (Plenum Press: New York, 1999).

21. U.S. EPA, Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001, U.S. Environmental Protection Agency, Washington, DC (1997).

22. U.S. EPA, Report of the Workshop on Selecting Input Distributions for Probabilistic Assessment, EPA/630/R-98/004, Washington, DC (1999).

23. Harris, C.M., "On Finite Mixtures of Geometric and Negative Binomial Distributions," *Commun, Statist.-Ther. Meth*. **12**, 987-1007 (1983).

24. Kanji, G.K., "A Mixture Model for Wind Shear Data," *J.Appl. Statist.* **12**, 49-58 (1985).

25. Wedel, M., Desarbo, W. S., Bult, J. R, Ramaswamy,V., "A Latent Class Poisson Regression Model for Heterogeneous Count Data," *Journal of Applied Econometrics*, Vol. 8, No. 4, 397-411 (1993).

26. Burmaster, D.E. and Wilson, A.M., "Fitted Second-order Finite Mixture Models to Data with Many Censored Values Using Maximum Likelihood Estimation," *Risk Analysis* **20**,

235-255 (2000).

27. Ott, W., "A Physical Explanation of the Log-normality of Pollutant Concentrations," *J. Of Air Waste and Management Association*, **40**: 1378-1383 (1990).

28. Titterington, D.M, Smith, A.F.M, and Makov, U.E., *Statistical Analysis of Finite Mixture Distributions*, (John Wiley & Sons, New York, NY, 1985).

29. Redner, R.A., Walker, H.F., "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, **26**, 195-239 (1984).

30. Everitt, B.S. and Hand, D.J., *Finite Mixture Distributions*, (Chapman & Hall, London, UK, 1981).

31. Cassie, R.M.," Some Uses of Probability Paper in the Analysis of Size Frequency Distributions," *Austral. J. Marine and Freshwater Res*., 5, pp. 513-523 (1954).

32. Casella, G., Berger, R.L., *Statistical Inference*, (Duxbury Press: Belmont, CA, 1990).

33. Dempster, A.P., Laird, N.M. and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM algorithm," *J. Royal Statist. Soc., Series B*, **39**,1-38 (1977).

34. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in FORTRAN*, (Cambridge University Press, New York, NY, 1992).

35. Efron, B. and Tibshirani, R.J., *an Introduction to the Bootstrap*, (Chapman & Hall, London, UK, 1993).

36. Angus, J.E., "Bootstrap One-sided Confidence Intervals for the Log-normal Mean," *Statistician*, **43**, 395-401(1994).

37. Freedman, D.A., and Peters, S.C., "Bootstrapping a Regression Equation: Some Empirical Results," *J. of the American Statistical Association*, **79**, 97-106 (1984).

38. Thombs, L.A., Schucany, W.R., "Bootstrap Predication Intervals for Autoregression," *J. of the American Statistical Association* **85**, 486-492 (1990).

39. Frey, H.C., Zheng, J., "Quantification of Variability and Uncertainty in Air Pollutant Emission Inventories: Method and Example Case Study for Utility NO_x Emissions, " *J. of Air & Waste Manage. Association,* **52**, 1083-1095 (2002).

40. Leroux, B., "Constant Estimation of a Mixing Distribution," *Annuals of Statistics*, 20:1350-1360 (1992).

Table 1. Selected Population Mixture Lognormal Distributions with Two Components

Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$	Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$	Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$
1	1.0	1.1	0.1	5	1.0	1.5	0.5	9	1.0	2.0	1.0
2	1.0	1.2	0.1	6	1.0	2.0	0.5	10	1.0	3.0	1.0
3	1.0	1.4	0.1	7	1.0	3.0	0.5	11	1.0	5.0	1.0
4	1.0	2.0	0.1	8	1.0	6.0	0.5	12	1.0	11.0	1.0

Table 2. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to the Sample from Mixture Populations with Varying Component Separation and Standard Deviation for n=100 and w=0.3

Population Parameters [a]				Fitted Dist. [b]	2.5 Percentile PV (CI) [c]	30 Percentile [d] PV (CI) [c]	50 Percentile PV (CI) [c]	75 Percentile PV (CI) [c]	97.5 Percentile PV (CI) [c]	Mean PV (CI) [c]	Standard Deviation PV (CI) [c]
μ_1	σ_1	μ_2	σ_2								
1.0	0.1	1.1	0.1	Mixture	0.87 (0.82-0.91)	1.01 (0.98-1.04)	1.06 (1.04-1.10)	1.14 (1.11-1.17)	1.28 (1.27-1.34)	1.06 (1.02-1.07)	0.11 (0.10-0.13)
				Single	0.87 (0.84-0.91)	1.01 (0.98-1.04)	1.06 (1.04-1.09)	1.14 (1.11-1.17)	1.28 (1.24-1.35)	1.06 (1.02 -1.07)	0.11 (0.09-1.13)
		1.2	0.1	Mixture	0.85 (0.82-0.94)	1.07 (1.03-1.11)	1.15 (1.11-1.18)	1.24 (1.20-1.27)	1.39 (1.33-1.44)	1.14 (1.11-1.16)	0.14 (0.12-0.15)
				Single	0.85 (0.85-0.94)	1.07 (1.03-1.09)	1.15 (1.10-1.16|)	1.24 (1.19-1.27)	1.39 (1.36-1.50)	1.14 (1.11-1.17)	0.14 (0.12-0.16)
		1.4	0.1	Mixture	0.87 (0.81-0.93)	1.07 (1.05-1.30)	1.34 (1.29-1.38)	1.43 (1.40-1.47)	1.58 (1.53-1.64)	1.27 (1.24-1.32)	0.21 (0.18-0.23)
				Single	0.87 (0.83-0.96)	1.07 (1.10-1.20)	1.34 (1.21-1.31)	1.43 (1.35-1.48)	1.58 (1.64-1.88)	1.27 (1.23-1.32)	0.21 (0.19-0.26)
		2.0	0.1	Mixture	0.87 (0.82-0.92)	1.77 (1.05-1.88)	1.94 (1.88-1.97)	2.04 (1.99-2.06)	2.18 (2.12-2.22)	1.71 (1.60-1.78)	0.46 (0.42-0.50)
				Single	0.87 (0.76-0.99)	1.77 (1.25-1.48)	1.94 (1.51-1.75)	2.04 (1.97-2.20)	2.18 (2.68-3.45)	1.71 (1.60-1.82)	0.46 (0.47 -0.67)
1.0	0.5	1.5	0.5	Mixture	0.44 (0.35-0.64)	1.01 (0.87-1.14)	1.27 (1.13-1.41)	1.63 (1.50-1.82)	2.62 (2.21-3.03)	1.34 (1.23-1.45)	0.55 (0.47-0.64)
				Single	0.44 (0.44-0.64)	1.01 (0.87-1.09)	1.27 (1.11-1.35)	1.63 (1.48-1.83)	2.62 (2.33-3.36)	1.34 (1.16-1.39)	0.55 (0.48-0.77)
		2.0	05	Mixture	0.43 (0.35-0.62)	1.34 (1.12-1.57)	1.69 (1.57-1.88)	2.13 (1.98-2.31)	3.01 (2.66-3.43)	1.67 (1.58-1.84)	0.68 (0.59-0.78)
				Single	0.43 (0.46-0.72)	1.34 (1.02-1.31)	1.69 (1.34-1.67)	2.13 (1.84-2.34)	3.01 (3.10-4.70)	1.67 (1.53-1.87)	0.68 (0.68-1.12)
		3.0	0.5	Mixture	0.46 (0.35-0.63)	1.75 (1.12-2.44)	2.67 (2.47-2.85)	3.13 (2.98-3.29)	4.02 (3.65-4.31)	2.35 (2.15-2.58)	1.06 (0.94-1.15)
				Single	0.46 (0.45-0.82)	1.75 (1.21-1.72)	2.67 (1.76-2.33)	3.13 (2.62-3.63)	4.02 (5.67-8.08)	2.35 (2.14-2.82)	1.06 (1.25-2.22)
		6.0	0.5	Mixture	0.45 (0.36-0.65)	1.88 (1.19-5.45)	5.64 (5.48-5.88)	6.13 (6.03-6.34)	6.93 (6.70-7.24)	4.40 (4.08-4.97)	2.37 (2.09-2.51)
				Single	0.45 (0.34-0.97)	1.88 (1.57-2.78)	5.64 (2.68-4.25)	6.13 (4.89-7.57)	6.93 (12.3-27.3)	4.40 (4.03-6.22)	2.37 (3.57-8.71)
1.0	1.0	2.0	1.0	Mixture	0.20 (0.15-0.40)	1.04 (0.84-1.29)	1.48 (1.30-1.76)	2.25 (1.93-2.60)	4.24 (3.42-5.23)	1.68 (1.51-1.91)	1.07 (0.85-1.26)
				Single	0.20 (0.21-0.46)	1.04 (0.74-1.13)	1.48 (1.12-1.60)	2.25 (1.85-2.65)	4.24 (3.99-7.99)	1.68 (1.47-2.09)	1.07 (1.05-2.12)
		3.0	1.0	Mixture	0.20 (0.14-0.37)	1.77 (1.05-2.08)	2.46 (2.07-2.68)	3.26 (2.90-3.55)	5.06 (4.39-6.09)	2.41 (2.12-2.65)	1.33 (1.14-1.53)
				Single	0.20 (0.22-0.60)	1.77 (0.94-1.57)	2.46 (1.54-2.32)	3.26 (2.65-4.01)	5.06 (6.16-12.7)	2.41 (2.16-3.19)	1.33 (1.70-3.77)
		5.0	1.0	Mixture	0.25 (0.12-0.33)	3.35 (1.01-3.97)	4.39 (3.95-4.72)	5.26 (4.95-5.62)	7.00 (6.39-7.72)	3.83 (3.37-4.22)	2.04 (1.88-2.35)
				Single	0.25 (0.23-0.82)	3.25 (1.24-2.37)	4.39 (2.22-3.78)	5.26 (4.25-6.93)	7.00 (10.8-26.5)	3.83 (3.51-5.57)	2.04 (3.12-9.16)
		11.0	1.0	Mixture	0.23 (0.14-0.37)	8.87 (1.05-9.87)	10.5 (9.85-10.8)	11.4 (11.0-11.7)	12.9 (12.4-13.5)	8.11 (6.94-8.77)	4.63 (4.26-5.06)
				Single	0.23 (0.25-1.17)	8.87 (1.88-4.74)	10.5 (4.01-8.29)	11.4 (9.14-17.6)	12.9 (30.9-96.7)	8.11 (8.05-17.0)	4.63 (10.5-43.5)

[a]: Arithmetic mean and standard deviation of each component.

[b]: Fitted mixture is a two component lognormal, single distribution is lognormal.

[c]: PV=Population value, CI =95 % confidence interval. Shading indicates that confidence interval does not enclose population value.

[d]: That the 30 percentile instead of 25 percentile was chosen is because there is weight of 0.3. The purpose is to observe how confidence interval varies at the inflection point.

Table 3. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to Sample from Mixture Populations with Varying Component Separation and Standard Deviation for n=100 and w=0.5

Population Parameters [a]				Fitted Dist. [b]	2.5 Percentile PV (CI) [c]	30 Percentile PV (CI) [c]	50 Percentile PV (CI) [c]	75 Percentile PV (CI) [c]	97.5 Percentile PV (CI) [c]	Mean PV (CI) [c]	Standard Deviation PV (CI) [c]
μ_1	σ_1	μ_2	σ_2								
1.0	0.1	1.1	0.1	Mixture	0.84 (0.81-0.89)	0.97 (0.94-1.00)	1.04 (1.01-1.07)	1.12 (1.09-1.16)	1.26 (1.22-1.32)	1.04 (1.02-1.07)	0.11(0.10 -0.13)
				Single	0.84 (0.81-0.88)	0.97 (0.94-0.99)	1.04 (1.01-1.07)	1.12 (1.09-1.15)	1.26 (1.22-1.33)	1.04 (1.02-1.07)	0.11(0.09 -0.13)
		1.2	0.1	Mixture	0.84 (0.80-0.89)	0.99 (0.95-1.03)	1.10 (1.05-1.14)	1.20 (1.16-1.24)	1.37 (1.31-1.42)	1.10 (1.07-1.12)	0.14(0.13 - 0.16)
				Single	0.84 (0.80-0.89)	0.99 (0.96-1.03)	1.10 (1.05-1.12)	1.20 (1.15-1.23)	1.37 (1.33-1.47)	1.10 (1.07-1.13)	0.14 (0.12 -0.16)
		1.4	0.1	Mixture	0.84 (0.80-0.89)	1.00 (0.96-1.04)	1.22 (1.08-1.32)	1.40 (1.34-1.43	1.56 (1.50-1.60)	1.20 (1.15-1.23)	0.22 (0.20 -0.24)
				Single	0.84 (0.76-0.87)	1.00 (0.99-1.09)	1.22 (1.13-1.23)	1.40 (1.28-1.41)	1.56 (1.58-1.83)	1.20 (1.16-1.24)	0.22 (0.20 -0.26)
		2.0	0.1	Mixture	0.85 (0.79-0.88)	0.99 (0.94-1.03)	1.20 (1.07-1.90)	2.00 (1.93-2.03)	2.18 (2.10-2.21)	1.49 (1.38-1.58)	0.51 (0.49-0.53)
				Single	0.85 (0.61-0.82)	0.99 (1.00-1.21)	1.20 (1.30-1.51)	2.00 (1.63-1.96)	2.18 (2.38-3.19)	1.49 (1.39-1.61)	0.51 (0.44-0.66)
1.0	0.5	1.5	0.5	Mixture	0.41 (0.33-0.53)	0.83 (0.73-0.99)	1.19 (1.06-1.35)	1.58 (1.44-1.79)	2.58 (2.18-3.12)	1.26 (1.16-1.39)	0.57 (0.48-0.69)
				Single	0.41 (0.36-0.57)	0.83 (0.73-0.94)	1.19 (1.02-1.27)	1.58 (1.38-1.75)	2.58 (2.24-3.33)	1.26 (1.16-1.39)	0.59 (0.49-0.77)
		2.0	05	Mixture	0.39 (0.34-0.53)	0.88 (0.72-1.12)	1.48 (1.25-1.66)	1.99 (1.75-2.19)	2.96 (2.55-3.44)	1.49 (1.37-1.62)	0.71 (0.61-0.81)
				Single	0.39 (0.36-0.58)	0.88 (0.78-1.07)	1.48 (1.16-1.50)	1.99 (1.64-2.17)	2.96 (2.88-4.69)	1.49 (1.36-1.69)	0.71 (0.67-1.13)
		3.0	0.5	Mixture	0.41 (0.31-0.52)	0.87 (0.72-1.11)	2.06 (1.24-2.63)	2.95 (2.69-3.15)	3.89 (3.53-4.23)	1.97 (1.74-2.34)	1.13 (0.99-1.23)
				Single	0.41 (0.30-0.59)	0.87 (0.82-1.84)	2.06 (1.37-1.91)	2.95 (2.15-2.98)	3.89 (4.45-7.99)	1.97 (1.74-2.35)	1.13 (1.15-2.16)
		6.0	0.5	Mixture	0.37 (0.31-0.52)	0.88 (0.73-1.12)	2.30 (1.31-5.58)	5.99 (5.75-6.19)	6.79 (6.57-7.16)	3.44 (3.05-4.00)	2.55 (2.44-2.64)
				Single	0.17 (0.16-0.62)	0.88 (0.79-1.69)	2.30 (1.75-3.10)	5.99 (3.50-5.90)	6.79 (9.54-23.6)	3.44 (2.91-4.91)	2.55 (2.81-5.43)
1.0	1.0	2.0	1.0	Mixture	0.18 (0.12-0.28)	0.71 (0.51-0.93)	1.31 (1.07-1.54)	2.08 (1.74-2.35)	4.32 (3.19-5.28)	1.54 (1.3-1.72)	1.11 (0.85-1.41)
				Single	0.18 (0.15-0.37)	0.71 (0.50-0.87)	1.31 (0.95-1.46)	2.08 (1.64-2.49)	4.32 (3.75-7.95)	1.54 (1.33-1.96)	1.11 (0.97-2.32)
		3.0	1.0	Mixture	0.18 (0.12-0.30)	0.71 (0.50-1.09)	1.89 (1.37-2.37)	2.93 (2.49-3.30)	4.86 (4.12-5.91)	1.97 (1.75-2.27)	1.37 (1.18-1.61)
				Single	0.18 (0.14-0.42)	0.71 (0.57-1.05)	1.89 (1.13-1.84)	2.93 (2.10-3.36)	4.86 (5.27-11.8)	1.97 (1.75-2.77)	1.37 (1.54-3.84)
		5.0	1.0	Mixture	0.16 (0.12-0.27)	0.71 (0.49-1.02)	3.55 (1.36-4.16)	5.01 (4.42-5.28)	6.78 (6.01-7.64)	3.06 (2.56-3.44)	2.26 (2.05-2.46)
				Single	0.16 (0.12-0.49)	0.71 (0.63-1.47)	3.55 (1.44-2.80)	5.01 (3.09-5.80)	6.78 (9.46-26.3)	3.06 (2.66-4.98)	2.26(2.99-10.3)
		11.0	1.0	Mixture	0.16 (0.11-0.27)	0.69 (0.48-1.06)	8.83 (1.38-10.2)	10.9 (10.5-11.4)	12.7 (12.3-13.5)	6.02 (5.20-7.22)	5.09 (4.88-5.32)
				Single	0.16 (0.12-0.63)	0.69 (0.90-2.48)	8.83 (2.42-5.51)	10.9 (5.72-13.3)	12.7 (21.1-81.9)	6.02 (5.22-13.2)	5.09 (7.65-43.3)

[a]: Arithmetic mean and standard deviation of each component.
[b]: Fitted mixture is a two component lognormal, single distribution is lognormal.
[c]: PV=Population value, CI =95 % confidence interval. Shading indicates that confidence interval does not enclose population value.
[d]: That the 30 percentile instead of 25 percentile was chosen is because there is weight of 0.3. The purpose is to observe how confidence interval varies at the inflection point.

Table 4. Uncertainty of estimated parameters of mixture lognormal distribution fitted to 12-month average NO_x emission data for tangential coal-fired furnace with low NO_x burners and overfire air

Parameter	Units	2.5^{th} Percentile	Mean	97.5^{th} Percentile
Weight		0.078	0.249	0.523
μ_1	g/GJ	5.560	5.910	6.223
σ_1	g/GJ	0.044	0.236	0.460
μ_2	g/GJ	6.219	6.267	6.318
σ_2	g/GJ	0.042	0.109	0.223

Table 5. Comparison of Selected Statistics of the 95 Percent Confidence Interval for the Mean Based Upon a Mixture Distribution and Three Single Component Parametric Distributions (B=500).

Distribution Type	Absolute Uncertainty			Relative Uncertainty*	
	2.5% [L, U][b]	Mean [L, U][b]	97.5% [L, U][b]	(-) %	(+)%
Mixture [a]	475 [468,483]	504 [501, 506]	532 [530,535]	-5.6	5.7
Normal	466 [463,468]	505 [504, 506]	545 [543,548]	-7.7	7.9
Lognormal	466 [464,469]	505 [504, 505]	546 [543,550]	-7.7	8.1
Weibull	467 [465,469]	506 [505, 506]	543 [542,545]	-7.7	7.3

*: Negative Random Error= (2.5^{th} Percentile –Mean)/Mean,
Positive Random Error=(97.5^{th} Percentile –Mean)/Mean

[a]: Two component mixture lognormal distributions

[L, U][b]: Lower bound and upper bound based upon the precision of the average of ten bootstrap simulation

Table 6. Comparison of Selected Statistics of the 95 Percent Confidence Interval for the 95% Percentile of Variability Based Upon a Mixture Distribution and Three Single Component Parametric Distributions (B=500).

Distribution Type	Absolute Uncertainty			Relative Uncertainty*	
	2.5% [L, U][b]	Mean [L, U][b]	97.5% [L, U][b]	(-) %	(+)%
Mixture [a]	581 [578,584]	638 [631, 645]	750 [739,762]	-8.9	17.6
Normal	635 [633,638]	701 [699, 702]	768 [765,772]	-9.4	9.6
Lognormal	627 [623,633]	713 [711, 714]	813 [808,819]	-12.1	14.0
Weibull	627 [624,631]	692 [691, 693]	778 [770,785]	-9.4	12.4

*: Negative Random Error= (2.5^{th} Percentile –Mean)/Mean,
Positive Random Error=(97.5^{th} Percentile –Mean)/Mean

[a]: Two component mixture lognormal distributions

[L, U][b]: Lower bound and upper bound based upon the precision of the average of ten bootstrap simulation

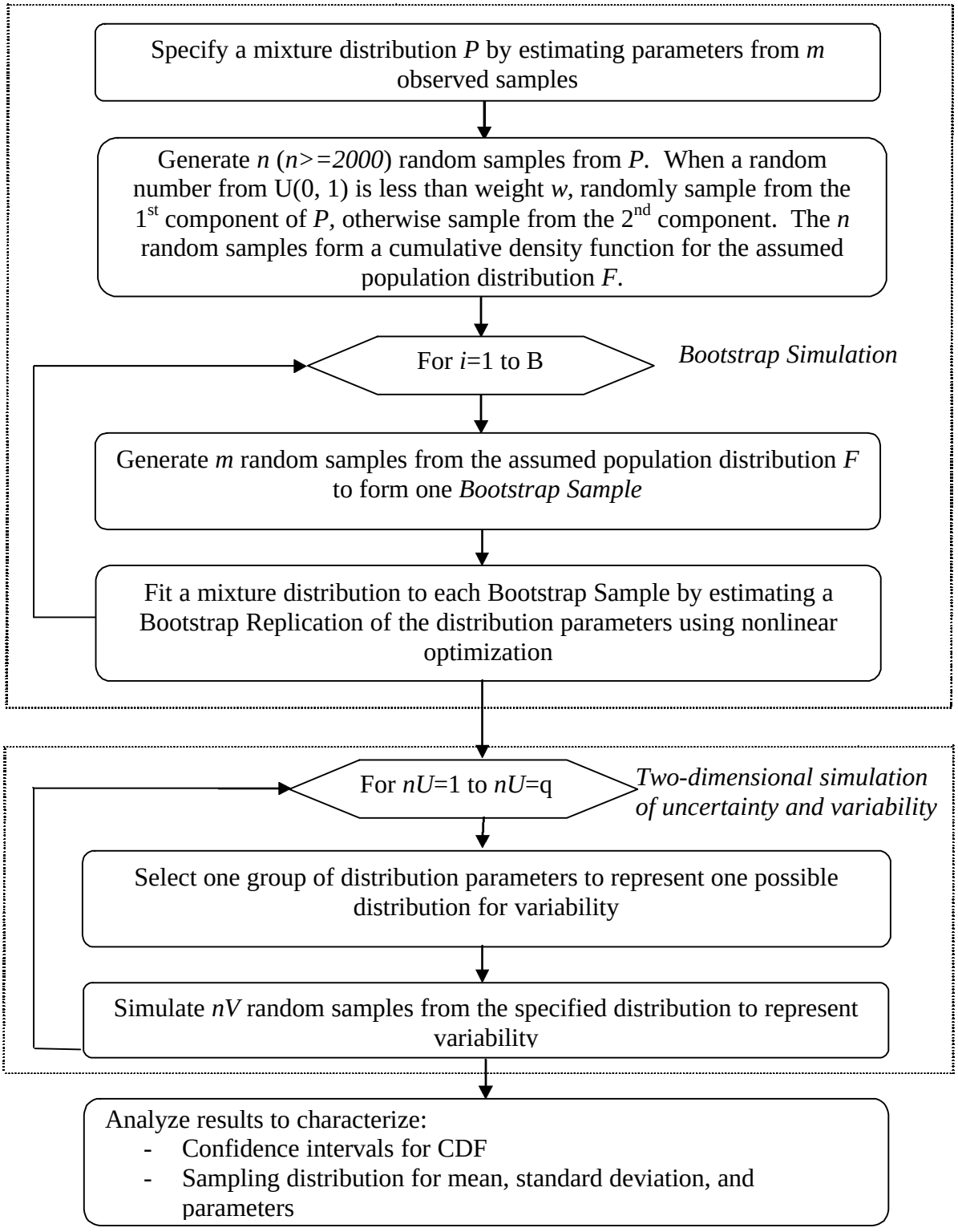


Figure 1. Simplified Flow Diagram for Quantification of Variability and Uncertainty Using Bootstrap Simulation Based upon Mixture Distributions

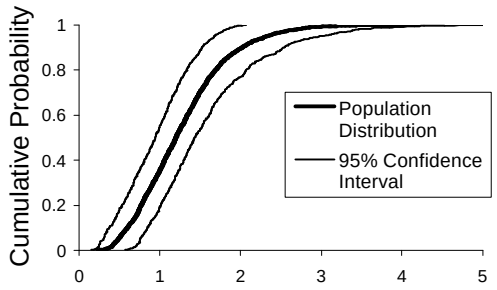
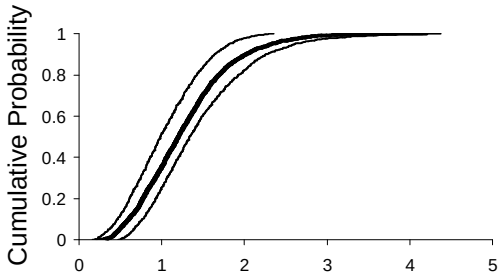
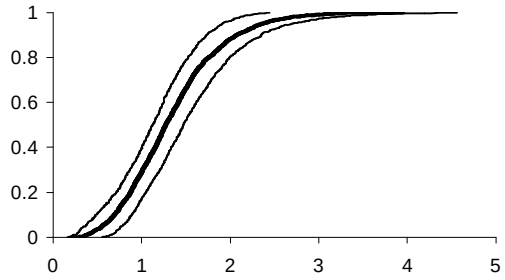
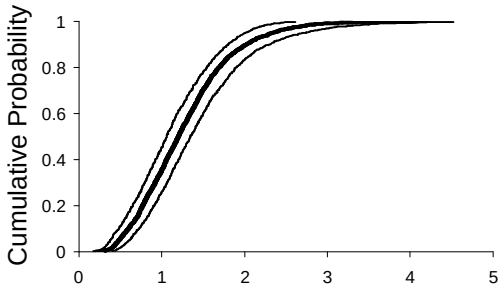
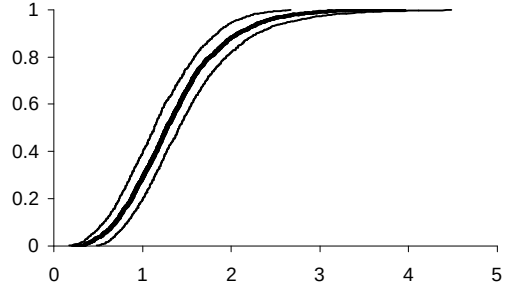
Sample Value	Weight=0.5	Weight=0.3	Weight=0.1
25		**Failed**	**Failed**
50			**Failed**
100			**Failed**

Figure 2. 95 Percent Confidence Intervals of Cumulative Distribution Functions of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=1.5, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)

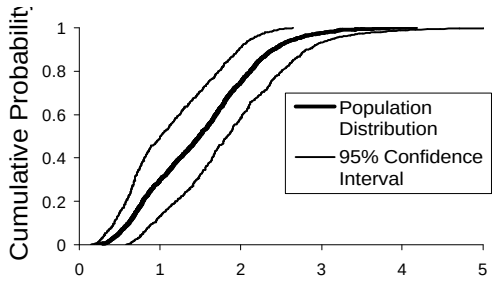
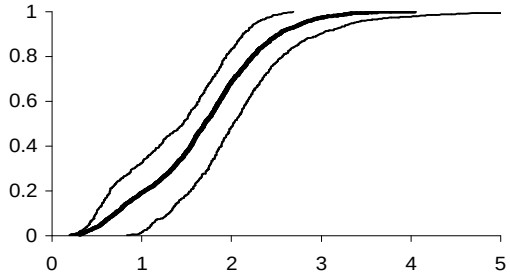
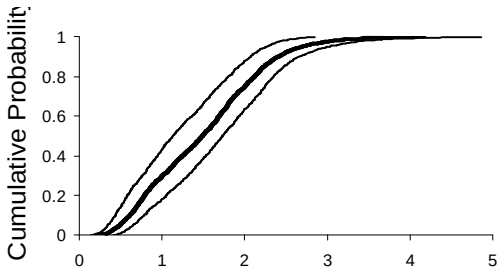
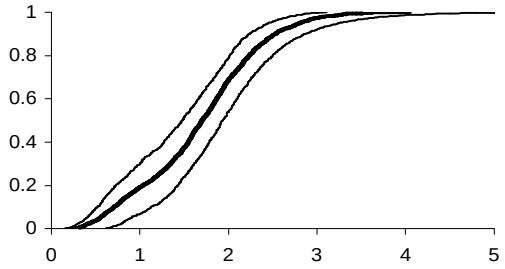
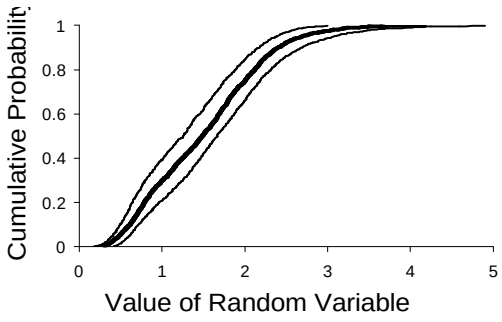
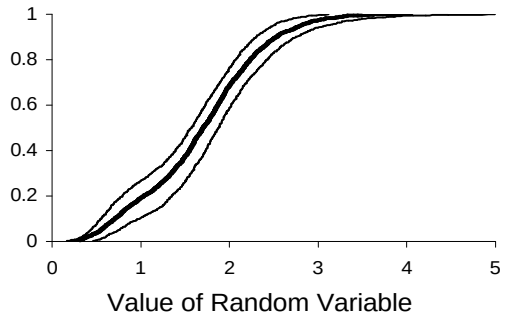
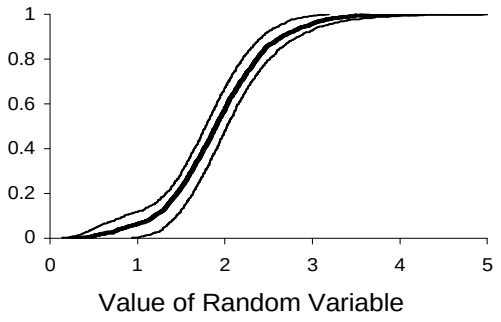
Sample Value	Weight=0.5	Weight=0.3	Weight=0.1
25			**Failed**
50			**Failed**
100			

Figure 3. 95 Percent Confidence Intervals of Cumulative Distribution Functions of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=2.0, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)

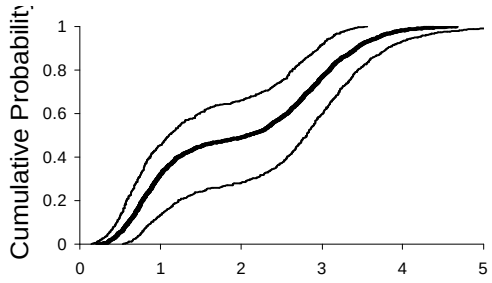
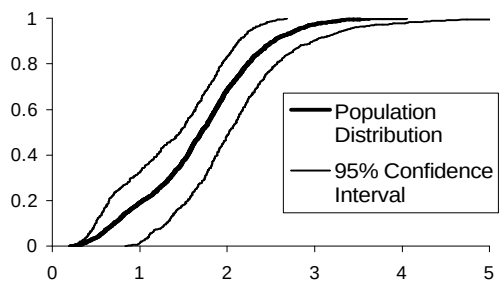
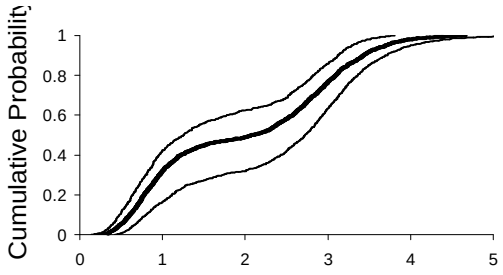
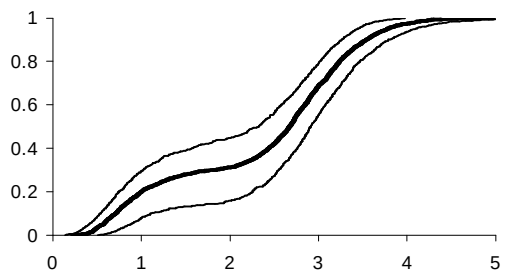
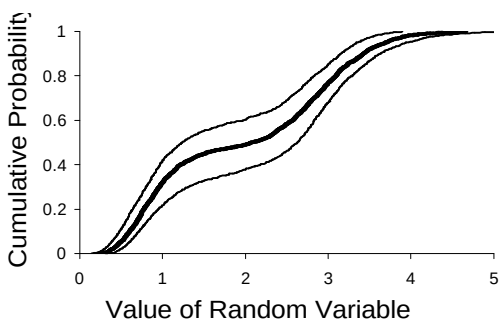
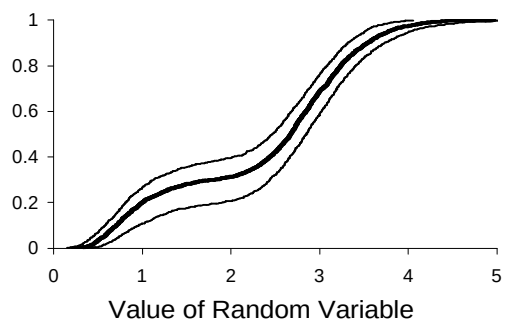
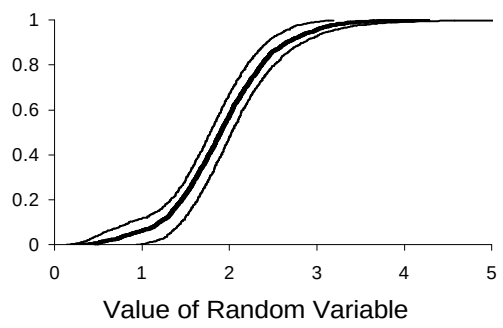
Sample Value	Weight=0.5	Weight=0.3	Weight=0.1
25			**Failed**
50			**Failed**
100			

Figure 4. 95 Percent Confidence Intervals of Cumulative Distribution Functions of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=3.0, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)

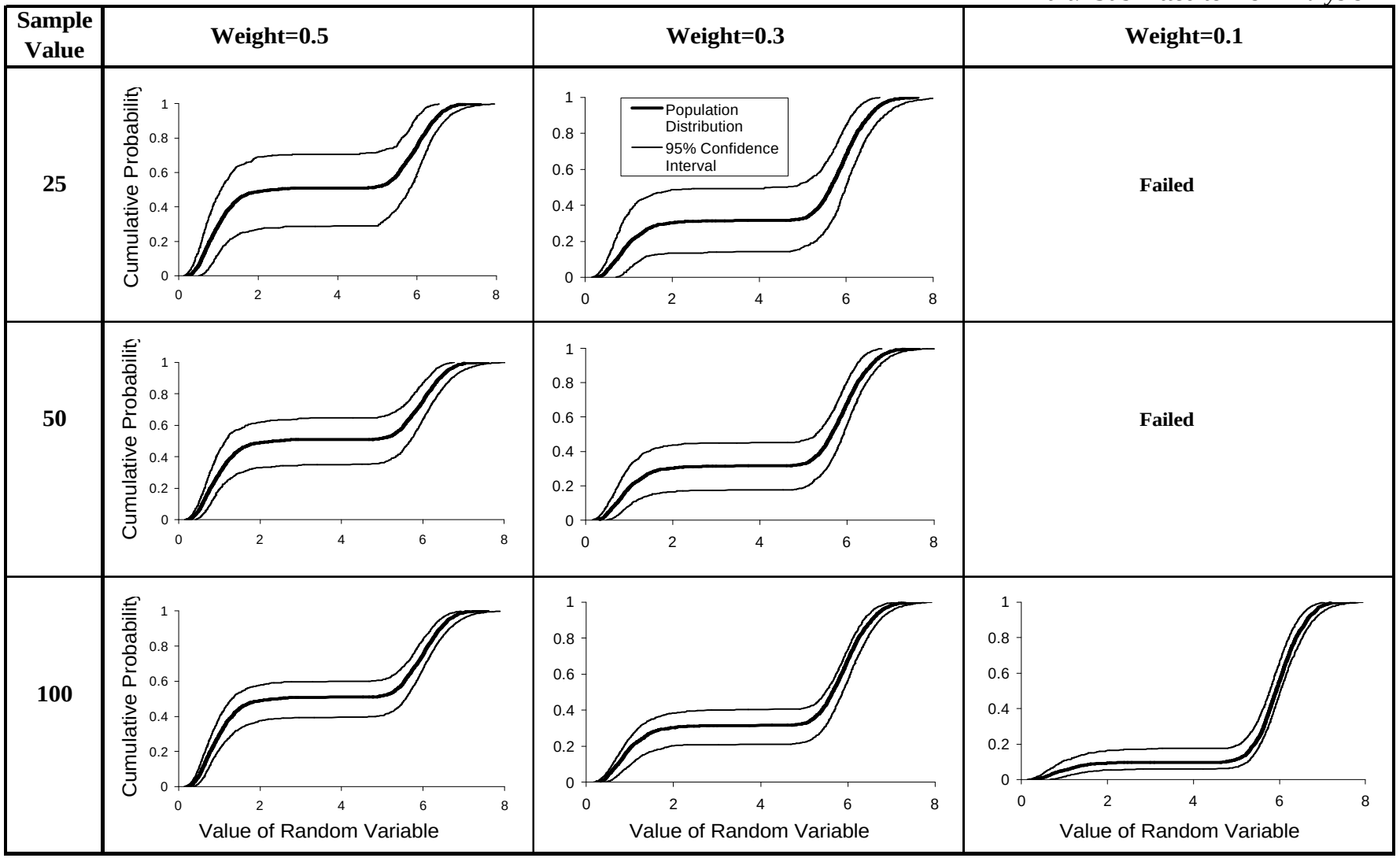


Figure 5. 95 Percent Confidence Intervals of Cumulative Distribution Functions of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=6.0, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)

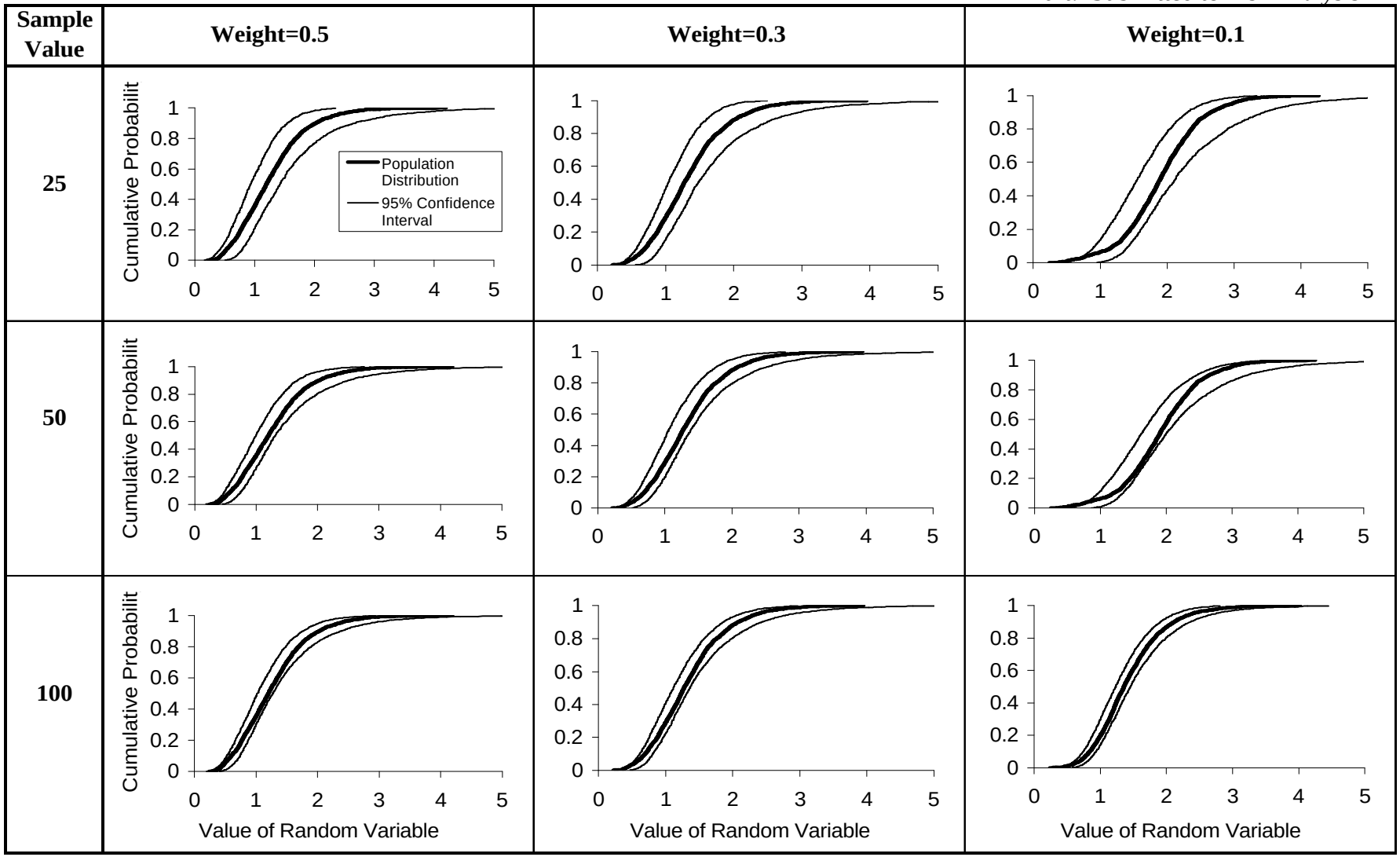


Figure 6. 95 Percent Confidence Intervals of Cumulative Distribution Functions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=1.5, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)

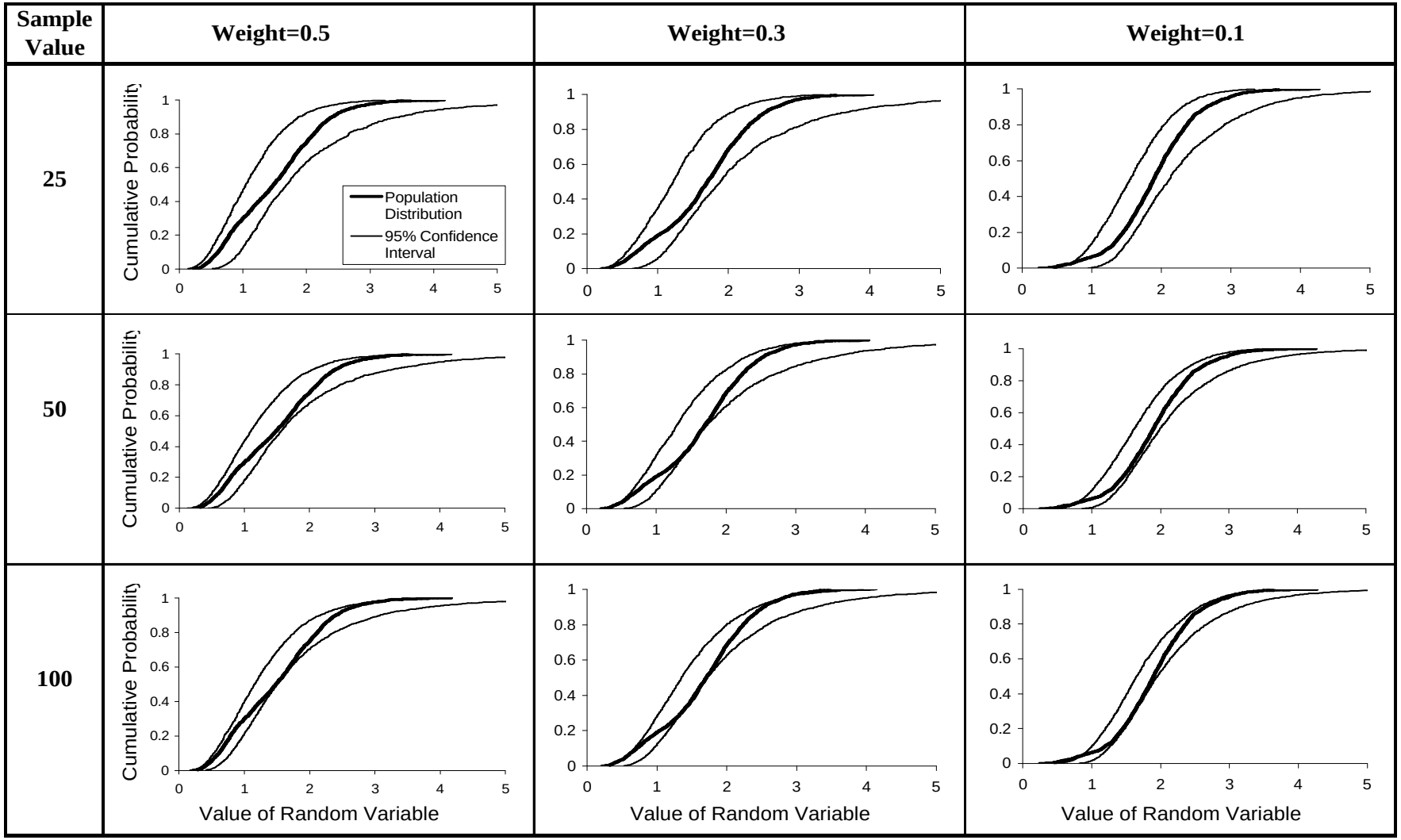
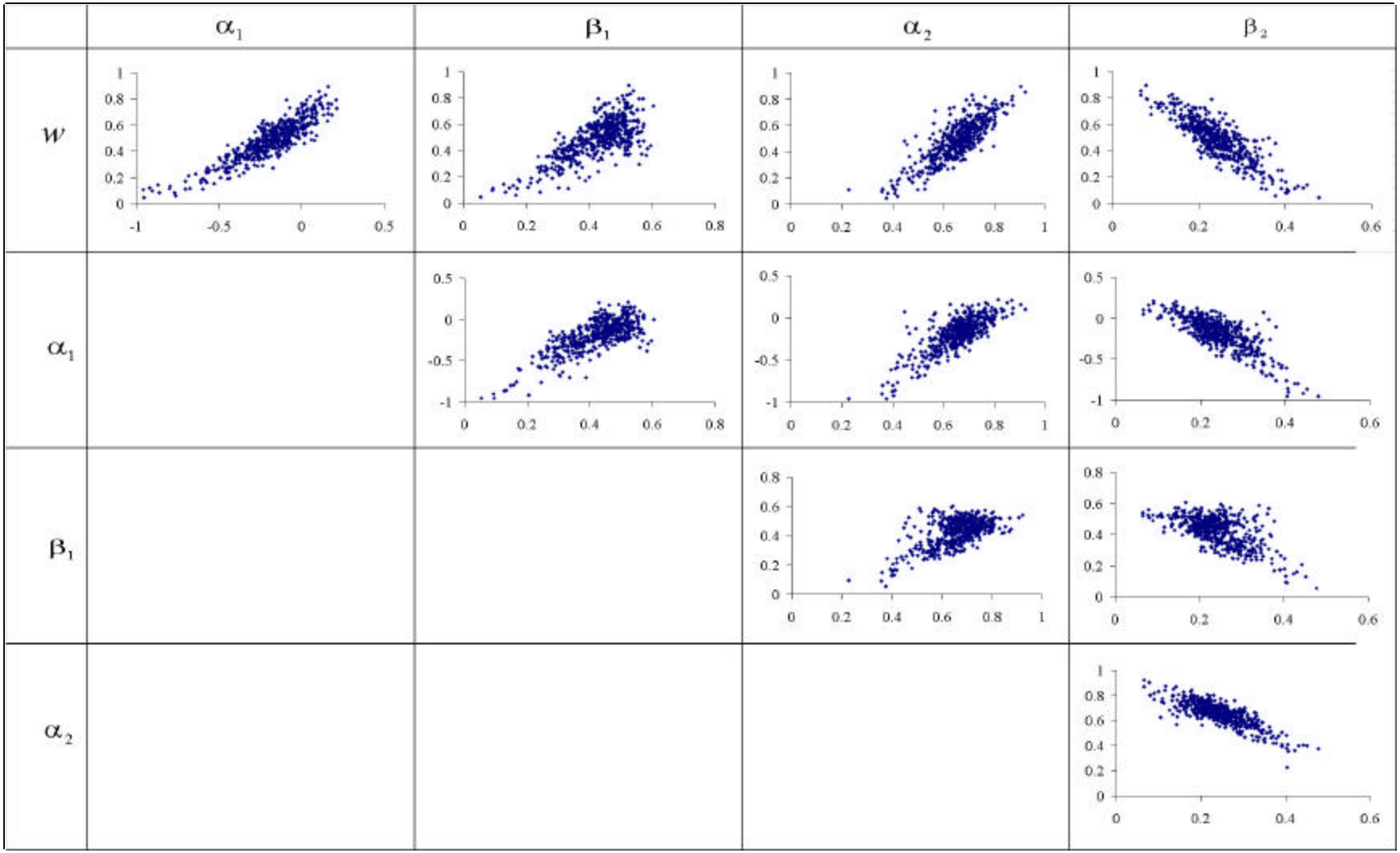
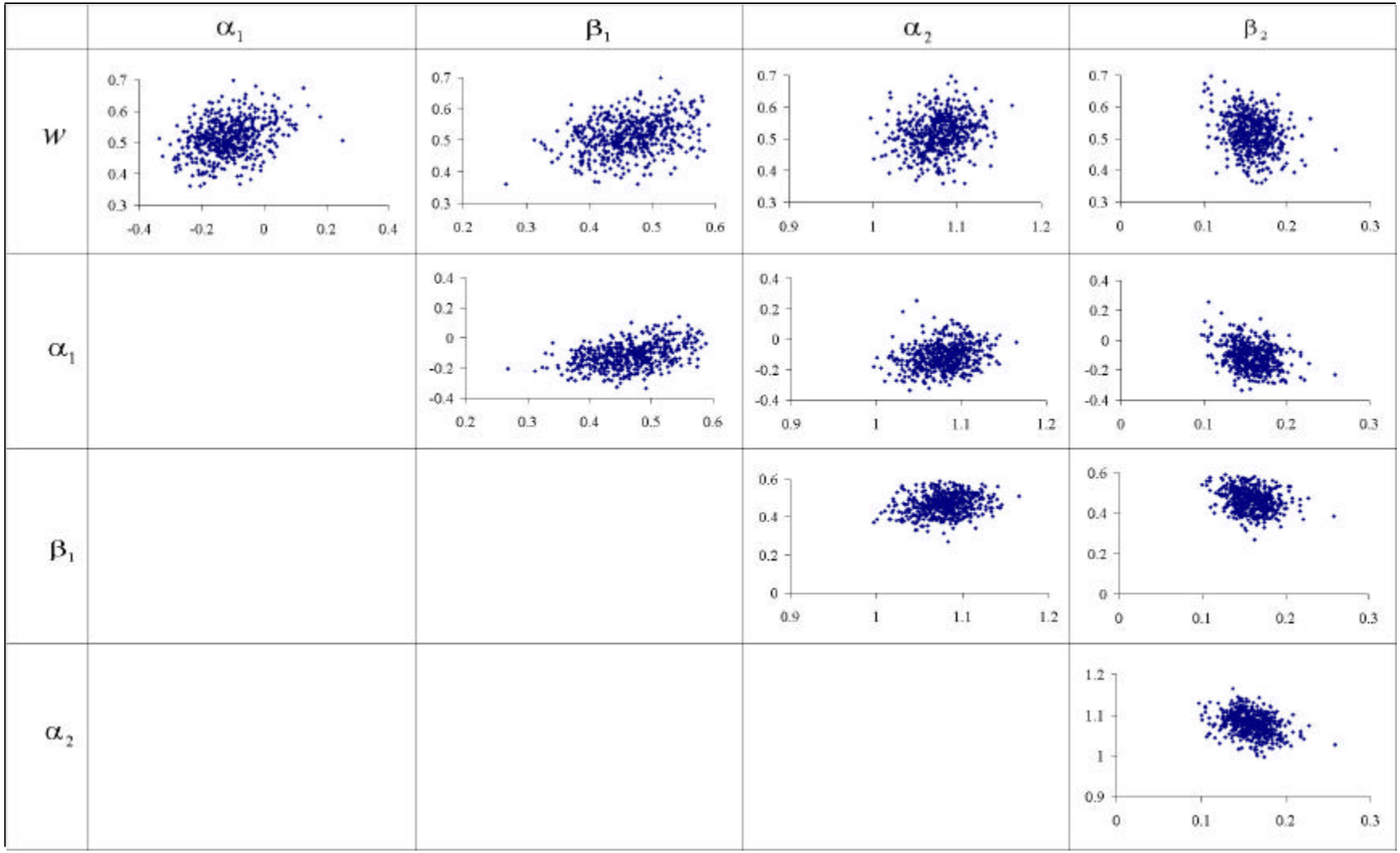


Figure 7. 95 Percent Confidence Intervals of Cumulative Distribution Functions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution (μ_1=1.0, σ_1=0.5, μ_2=2.0, σ_2=0.5) for n=25, 50 and 100, for w=0.1, 0.3 and 0.5 Based on Bootstrap Simulation (B=500)



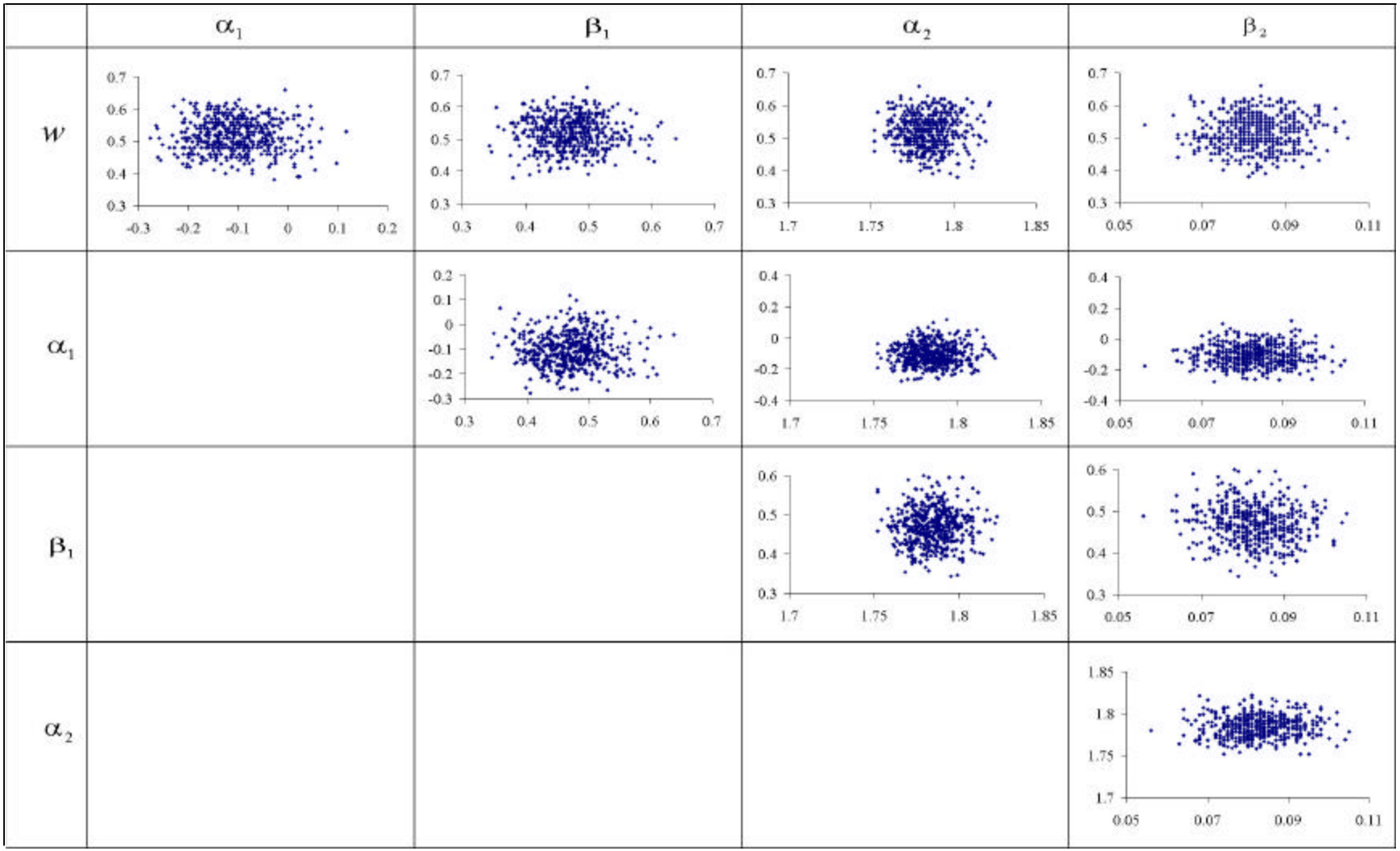
Note: For definitions of w,α_1, β_1, α_1, β_1, refers to the Equations (2) and (3) in text.

Figure 8. Scatter Plots of Bootstrap Simulation (B=500) Results for Parameters of Two Component Lognormal Mixture Distributions for n=100, w=0.5 with Slightly Separated Components (μ_1=1.0, σ_1=0.5, μ_2=2.0, σ_2=0.5).



Note: For definitions of w,α_1, β_1, α_1, β_1, refers to the Equations (2) and (3) in text.

Figure 9. Scatter Plots of Bootstrap Simulation (B=500) Results for Parameters of Two Component Lognormal Mixture Distributions for n=100, w=0.5 with Moderately Separated Components (μ_1=1.0, σ_1=0.5, μ_2=3.0, σ_2=0.5).



Note: For definitions of w,α_1, β_1, α_1, β_1, refers to the Equations (2) and (3) in text.

Figure 10. Scatter Plots of Bootstrap Simulation (B=500) Results for Parameters of Two Component Lognormal Mixture Distributions for n=100, w=0.5 with Highly Separated Components (μ_1=1.0, σ_1=0.5, μ_2=6.0, σ_2=0.

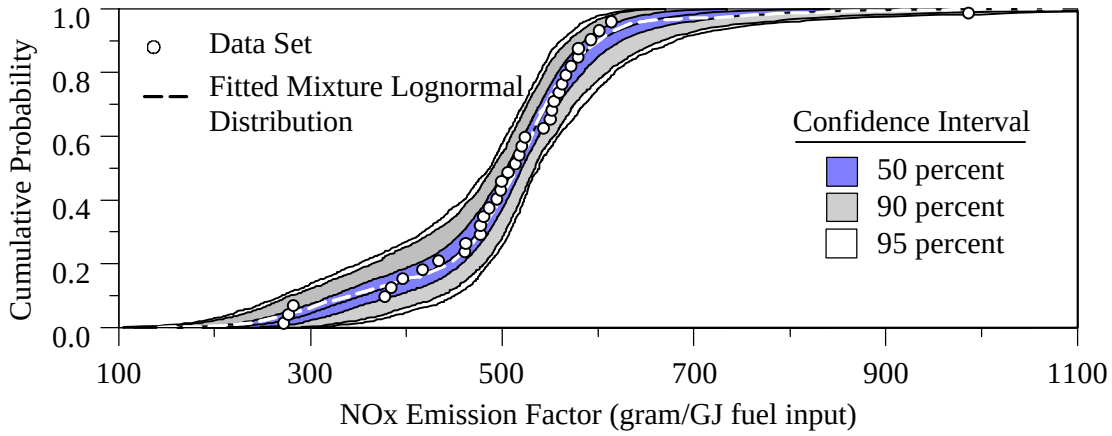


Figure 11. Probability band for fitted mixture lognormal distribution (n=36, B=500).

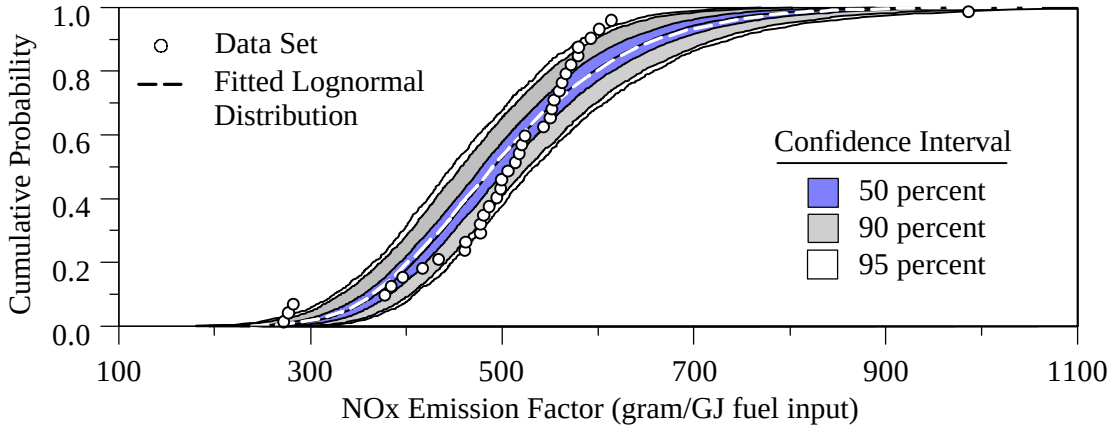


Figure 12. Probability band for fitted lognormal distribution (n=36, B=500).