Bootstrap Simulation and Two-Dimensional Monte Carlo Simulation: Dealing with Variability and Uncertainty, Mixture Distributions, Measurement Error, and Censored Data

H. Christopher Frey, Ph.D
Junyu (Allen) Zheng, Ph.D

NC STATE UNIVERSITY

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

Prepared for:
Society for Risk Analysis Annual Meeting
Sunday, December 8, 2001
New Orleans, LA

NC STATE UNIVERSITY

Workshop Materials

- Handouts
  - A hard copy of slides
  - A hard copy of AuvTool's user guide and technical report
  - A hard copy of the paper about mixture distributions submitted to *Risk Analysis*
- CD disk
  - AuvTool 98/ME version installation package
  - AuvTool 2000/XP version installation package
  - A PDF file of AuvTool's user guide
  - A PDF file of AuvTool's technical report

NC STATE UNIVERSITY

Agenda and Schedule

1:00 - 1:05	Welcome and Introduction Materials
1:05 - 1:35	Quantification of Variability
1:35 - 2:30	Quantification of Uncertainty in a Single Component Distribution
2:30 - 3:00	Introduction to AuvTool, Installation and its Use
3:00 - 3:15	Break
3:15 - 3:45	Censored Data
3:45 - 4:15	Mixture Distributions
4:15 - 4:45	Measurement Error
4:45 - 5:00	Summarization, Discussion and Evaluation
5:00 - 6:00 (Optional)	Demonstration of AuvTool and Questions

NC STATE UNIVERSITY

Outline

- Introduction
- Quantification of variability
- Quantification of uncertainty in a single component distribution
- Introduction to AuvTool
- Quantification of variability and uncertainty in censored datasets
- Characterization of variability and uncertainty based upon mixture distributions
- Characterization of variability and uncertainty with known measurement error

NC STATE UNIVERSITY

Introduction

- Limitations of qualitative or deterministic methods
- Increasing demand for quantitative analysis of variability and uncertainty in risk assessment, exposure assessment and emission estimation

NC STATE UNIVERSITY

Variability & Uncertainty

- Variability
  - Heterogeneity of values with respect to time, space, or a population
  - e.g, variation in feedstock or compositions; inter-plant variability in design, operation, and maintenance; and intra-plant variability
- Uncertainty
  - lack of knowledge regarding the true value of a quantity
  - e.g, statistical sampling error, measurement errors, and systematic errors

NC STATE UNIVERSITY

Instructional Objectives

- To identify general approaches for fitting distributions to data
- To describe and compare parameter estimation methods
- To briefly describe, and compare selected goodness-of-fit techniques
- To describe, calculate, and characterize confidence intervals for statistics
- To describe bootstrap simulation and apply it to characterize confidence intervals for fitted distributions
- To introduce two-dimensional Monte Carlo simulation for simultaneously characterizing variability and uncertainty
- To deal with special cases such as censored datasets,mixtures, and measurement error

NC STATE UNIVERSITY

Fitting Distributions to Data Sets

- Empirical (Non-Parametric) Approaches
- Parametric Approaches
  - Selection of parametric distributions
  - Selection of parameter estimation methods
- Evaluation of Goodness-of-Fit

NC STATE UNIVERSITY

Key Assumptions in Fitting Distributions to Data

- Random Sample
- Representative Sample

NC STATE UNIVERSITY

Statistical Estimation

- Inferences are made from samples of data regarding the characteristics of the population from which the data are a sample
- A "statistic" is a quantity that is estimated as a function of a random sample of data
- Examples of statistics include:
  - moments or central moments
  - percentiles or fractiles
  - parameters of distributions

NC STATE UNIVERSITY

Mean

- Mean (also Arithmetic Average, Average, Expected Value)
- For a continuous distribution:

$$E(x) = \int x\, f(x)\, dx$$

- For a data set:

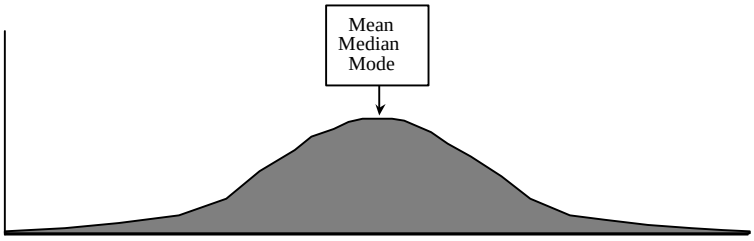
$$E(x) = \sum_{i=1}^{n} x_i p_i$$

- For equally weighted data:

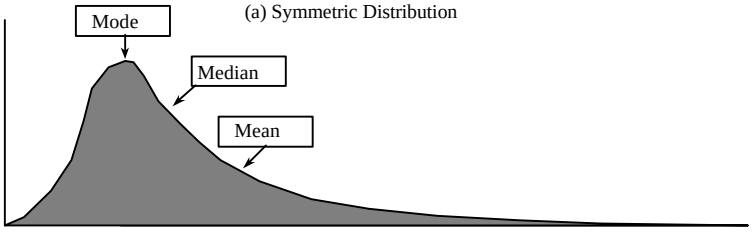
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

NC STATE UNIVERSITY

Comparison of Mean, Median, and Mode



(a) Symmetric Distribution



(b) Asymmetric Distribution

NC STATE UNIVERSITY

Variance

- Variance is the second central moment with respect to the mean:

$$\sigma^2 = \mu_2 = E\left[x - \mu_1\right]^2 = \int \left(x - \mu_1\right)^2 f(x)\, dx$$

- The variance may be estimated from a data set as follows:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

NC STATE UNIVERSITY

Coefficient of Variation

- The coefficient of variation is the standard deviation divided by the mean:

$$\nu = \frac{\sigma}{\mu}$$

- Also referred to as “relative standard deviation”
- Non-dimensional indication of the relative dispersion or width of a distribution

NC STATE UNIVERSITY

Third Central Moment and Skewness

- The third central moment is the basis for estimating skewness
- The third central moment is:

$$\mu_3 = E\left[x - \mu_1\right]^3 = \int \left(x - \mu_1\right)^3 f(x)\, dx$$

- The third central moment may be estimated as

$$m_3 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3}{n}$$

- The skewness is given by:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

NC STATE UNIVERSITY

Fourth Central Moment and Kurtosis

- Kurtosis is a measure of the "peakedness" or flatness of a distribution
- Larger kurtosis implies "pointier peaks"
- Kurtosis is based upon the fourth central moment:

$$\mu_4 = E\left[x - \mu_1\right]^4 = \int \left(x - \mu_1\right)^4 f(x)\, dx$$

- The fourth central moment may be estimated by

$$m_4 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^4}{n}$$

- Kurtosis is defined as:

$$\gamma_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{\mu_4}{\sigma^4}$$

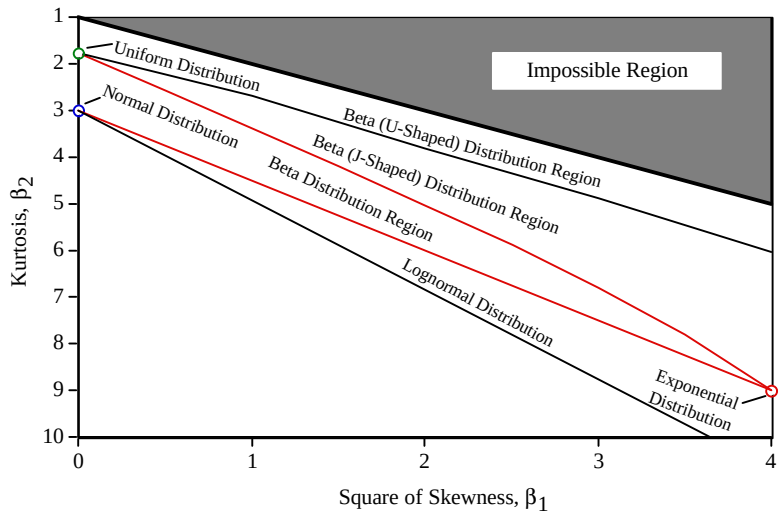
NC STATE UNIVERSITY

Selecting a Parametric Distribution to Fit to a Data Set

- The skewness and kurtosis of a data set can be used to help select a parametric distribution with similar shape
- This is an empirical approach to selecting a parametric distribution
- This approach may not be the most appropriate one to use
- Physical constraints and processes that generate data and distributions should also be considered

NC STATE UNIVERSITY

Empirical Basis for Selecting a Parametric Distribution: Moment Plane



NC STATE UNIVERSITY

Theoretic or Practical Basis for Common Parametric Distributions

- Normal Distribution: Asymptotic for central limit theorem for sums. Useful for random measurement errors and physical quantities when coefficient of variation is small
- Lognormal Distribution: Asymptotic for central limit theorem for products. Mixing processes. Useful for non-negative quantities that vary by orders-of-magnitude
- Weibull, Gamma: alternatives to Lognormal for non-negative quantities; different weighting of tails
- Beta: Useful for bounded quantities (e.g., 0 to 1), and for representing expert judgments
- Uniform, Triangular: Useful for representing expert judgment
- Spline, Empirical: Useful for representing data; often used to digitize expert judgments

NC STATE UNIVERSITY

Empirical Versus Parametric Distributions

- Both approaches are based on assumption of a random, representative data set
- A strictly empirical approach does not involve extrapolation beyond the range of observed data
- Artifacts of the shape of an empirical distribution may be attributable to random fluctuations

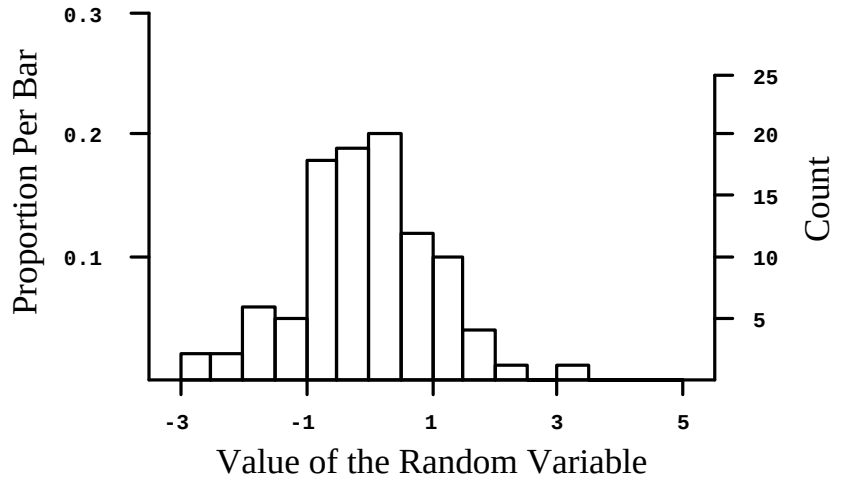
NC STATE UNIVERSITY

Visualization of Data

- Visualization of data is a useful and important means for gaining insight into the characteristics of the data
  - central tendency
  - dispersion
  - skewness
  - kurtosis

NC STATE UNIVERSITY

Visualizing Data: Histogram



NC STATE UNIVERSITY

Visualizing Data: Cumulative Distribution Functions

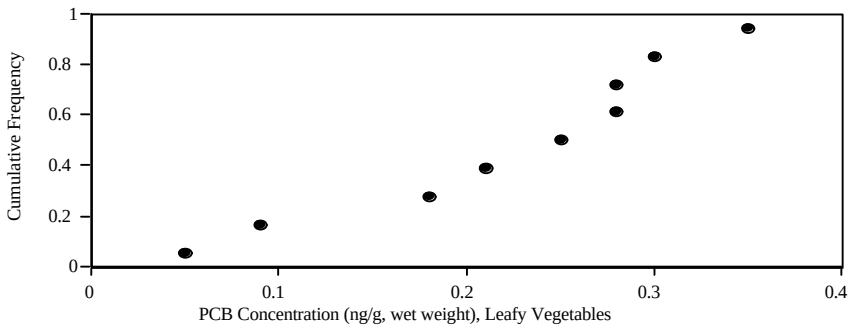
- Cumulative distribution functions are a quantitative way to represent empirical distributions of data
- The Hazen plotting position is often used:

$$F_X(x_i) = \Pr(X < x_i) = \frac{i - 0.5}{n}, \text{ for } i = 1, 2, ..., n \text{ and } x_1 < x_2 < ... < x_n$$

- The general approach is to
  - rank order the data in ascending order
  - assign a rank, i, to each data point (from 1 to n)
  - calculate the estimated cumulative probability
  - Plot cumulative probability versus x

NC STATE UNIVERSITY

An Example of Empirical Distribution



An empirical distribution is defined as a discrete distribution, *F*, that gives equal probability, 1/n, to each value x_i in the dataset, **x** (Efron, 1979).

NC STATE UNIVERSITY

Variability and Uncertainty

- Typically, data sets represent *variability* in a quantity over time, space, or members of a population
- Data are typically a sample from a population
- Ideally, data are a random and representative sample
- Can make inferences regarding estimated population statistics and distribution
- Lack of knowledge regarding the true population characteristics.
- Lack of knowledge = *uncertainty*

NC STATE UNIVERSITY

Uncertainty and Sampling Distributions

- Uncertainty due to small sample size
  - Random fluctuations due to "sampling error"
  - Quantified using confidence intervals
- Any statistic of a random variable is itself a random variable (e.g., mean, variance)
- The probability distribution for a statistic is referred to as a "sampling distribution."
- Important for evaluating whether your distribution model reasonably represents the data
- Can be calculated various ways, for example:
  - Analytical solutions (restricted situations)
  - Numerical methods (more generally applicable)

NC STATE UNIVERSITY

Sampling Distributions and Confidence Intervals

- A sampling distribution is the basis for a confidence interval
- A confidence interval is based upon specified percentiles of a sampling distribution
  - For example, a 95 percent confidence interval is typically enclosed by the 2.5th and 97.5th percentiles of the sampling distribution
- In principle, sampling distributions and confidence intervals can be developed for any statistic

NC STATE UNIVERSITY

Confidence Intervals for the Mean

- Confidence intervals for means are often estimated based upon a normality assumption
- This assumption may be invalid for small data sets and/or highly skewed data sets
- We review the conventional analytical approach to confidence intervals for the mean
- We present a numerical method for estimating confidence intervals based upon bootstrap simulation

NC STATE UNIVERSITY

Confidence Interval for the Mean

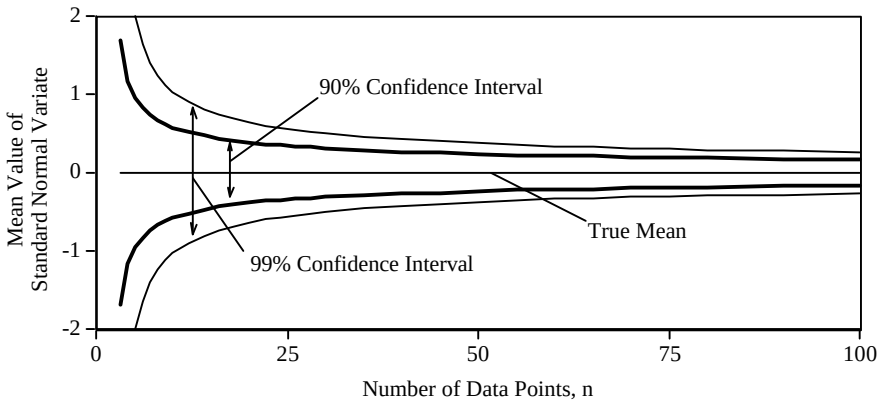
- The confidence interval for the mean is based upon the standard error of the mean and a standardized distribution:

$$\bar{x} \pm c \frac{s}{\sqrt{n}}$$

- The standardized distribution is often assumed to be either the student-t or normal distribution
- For n > 30, there is not much difference
- The student-t distribution is wider for small n

NC STATE UNIVERSITY

Effect of Sample Size on Confidence Intervals for Mean



NC STATE UNIVERSITY

Confidence Interval for the Variance

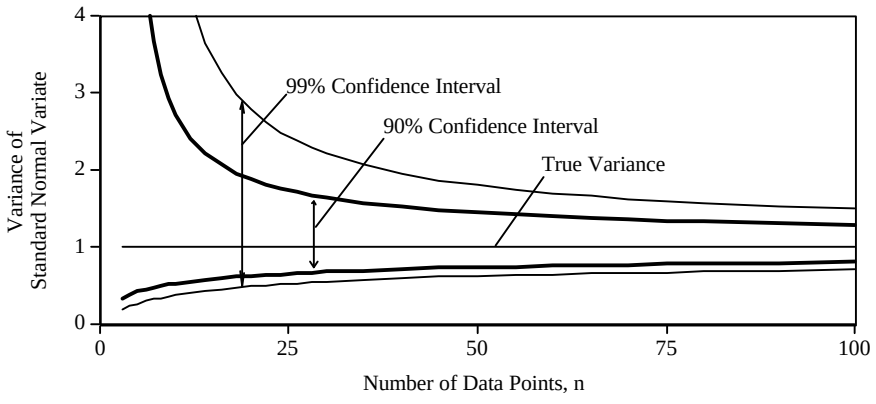
- The confidence interval for the variance can be estimated analytically for a normally distribution population:

$$\Pr\left(c_{\alpha/2,n-1} \le \frac{(n-1)\,s^2}{\sigma^2} \le c_{1-\alpha/2,n-1}\right) = 1-\alpha$$

- The standardized distribution used here is the chi-square distribution

NC STATE UNIVERSITY

Confidence Interval for the Variance



NC STATE UNIVERSITY

Implications of Sampling Error and Sampling Distributions

- Any statistic that you calculate from a data set is only one estimate of the true population value of that statistic
- In order to evaluate the adequacy of a fitted distribution, you should consider the range of possible values for statistics, such as the parameters of the fitted distribution
- Analytical solutions work in only a few cases
- A numerical method is more broadly applicable

NC STATE UNIVERSITY

Analytical Method: Advantage and Disadvantage

- Advantage
  - Can get an exact estimate of confidence interval
  - Simple to calculate
- Disadvantage
  - Confidence intervals for means are often estimated based upon a normality assumption
  - This assumption may be invalid for small data sets and/or highly skewed data sets
  - Analytical solutions work in only a few cases
  - Can not calculate confidence intervals for some statistics, e.g., parameters in a distribution

NC STATE UNIVERSITY

Numerical Method: Bootstrap Simulation

- Introduced by Efron in 1979
- A means for calculating confidence intervals for statistics in a general manner for situations in which analytical solutions are not available

NC STATE UNIVERSITY

Bootstrap Simulation: Resampling at Random from a Data Set

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

$$\hat{F} \rightarrow (x_1^*, x_2^*, \ldots, x_n^*)$$

Example Bootstrap Samples

$$\mathbf{x}^{*1} = (x_3, x_5, x_1, x_5, x_2)$$
$$\mathbf{x}^{*2} = (x_4, x_1, x_2, x_4, x_4)$$
$$\mathbf{x}^{*3} = (x_2, x_4, x_3, x_3, x_2)$$

NC STATE UNIVERSITY

Bootstrap Replications

- **For each bootstrap sample, calculate (replicate) a statistic:**

$$\hat{\theta}^{*} = s(\mathbf{x}^{*})$$

- **Repeat the replications *B* times:**

$$\hat{\theta}^{*}_{b} = s(\mathbf{x}^{*b})$$

- **b = 1, *B***

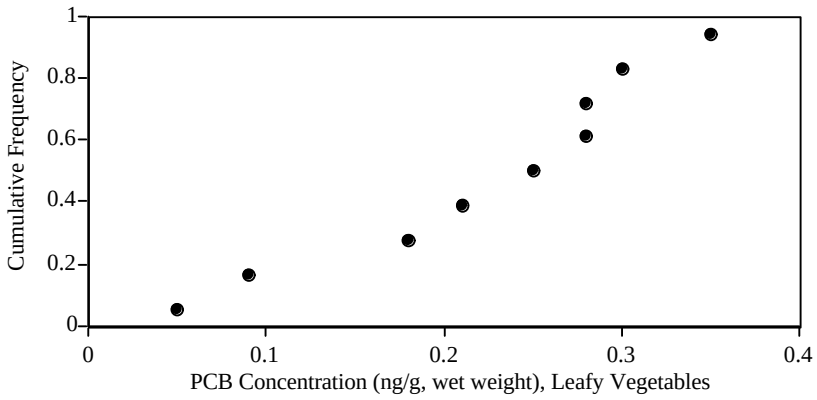
NC STATE UNIVERSITY

Bootstrap Simulation

- Original data set with *n* values
- *B* bootstrap samples of the data set, drawn from a distribution *F*
  - Resampling
  - Parametric distribution
- *B* replications of statistic of interest
  - Confidence intervals
  - Sampling distributions

NC STATE UNIVERSITY

Bootstrap Simulation: Example of the Leafy Vegetable Data Set



NC STATE UNIVERSITY

Criteria for Selecting a Parameter Estimation Method

- Consistency: Converges to the "true" value of the parameter as the number of samples increases.
- Lack of Bias: Average value of the parameter estimate that is equal to that of the population value.
- Efficiency: Minimum variance in the sampling distribution of the estimate.
- Sufficiency: Makes maximum use of information contained in a data set.
- Robustness: Works well even if there are departures from the underlying distribution.
- Practicality: Computationally efficient

NC STATE UNIVERSITY

Fitting Distribution to Data Parameter Estimation Methods

- Method of Matching Moments:
  - Typically involves matching the mean and variance of the distribution to the mean and variance of the data set (for a 2 parameter distribution)
  - In general involves matching *m* moments or central moments of the distribution to those of the data, where *m* = number of parameters.
  - Example: parameters of the Normal distribution are the mean and standard deviation of the data

NC STATE UNIVERSITY

Fitting Distribution to Data: Parameter Estimation Methods

- Maximum Likelihood Estimation
  - Select distribution parameters so that the fitted distribution is the one most likely to produce the observed data set
  - Involves maximization of the likelihood function:

$$L(\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, ..., \theta_k)$$

  - The log-likelihood function is often more convenient to use/program, since it is written as a sum rather than as a product

NC STATE UNIVERSITY

Examples of Log-Likelihood Functions

Name of Distribution [a]	Log-likelihood Function
Normal (μ = mean, σ = standard deviation)	$J(m,s) = -n\ln s - \frac{n}{2}\ln(2p) - \sum_{i=1}^{n}\left\{\frac{(x_i - m)^2}{2s^2}\right\}$
Lognormal (μ = mean, σ = standard deviation, of log-transformed data)	$J(m,s) = -n\ln s - \frac{n}{2}\ln(2p) - \sum_{i=1}^{n}\left\{\frac{(\ln(x_i) - m)^2}{2s^2}\right\}$
Gamma (a = shape, b = scale, parameters)	$J(a,b) = -n\{a\ln(b) + \ln[\Gamma(a)]\} + \sum_{i=1}^{n}\left\{(a-1)\ln(x_i) - \frac{x_i}{b}\right\}$
Weibull (a = shape, b = scale, parameters)	$J(a,b) = -n\ln\left(\frac{a}{b}\right) + \sum_{i=1}^{n}\left\{(a-1)\ln\left(\frac{x_i}{b}\right) - \left(\frac{x_i}{b}\right)^a\right\}$
Beta (a = shape, b = scale, parameters)	$J(a,b) = -n\ln\left\{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\right\} + \sum_{i=1}^{n}\{(a-1)\ln(x_i) - (b-1)\ln(1-x_i)\}$

[a] Note: Parameter values are different for each type of distribution even though the same symbol may be used to represent parameters of different distributions.

NC STATE UNIVERSITY

Which Parameter Estimation Method Should Be Used

- No method is necessarily always best
- Sometimes one method will fail for a particular data set
- MLE is often considered a more efficient method
- The specific values of distribution parameters will be different for a given data set if a different estimation method is used

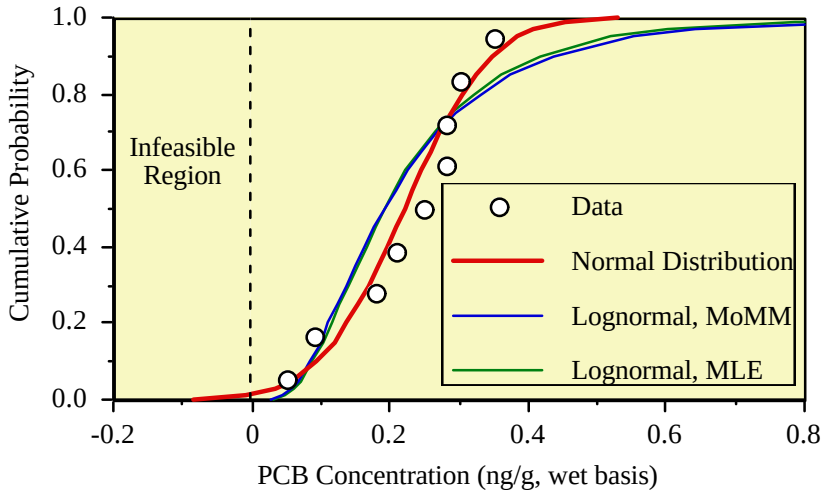
NC STATE UNIVERSITY

Parameter Estimation Using Probability Plots

- Probability plotting is most appropriate as a "goodness-of-fit" technique
- It is often not the most satisfactory method for estimating parameters
- To create a probability plot, data must be rank ordered
- Least-squares regression techniques are based on an assumption of statistical independence of data
- This assumption is violated in probability plots

NC STATE UNIVERSITY

Fitting Distributions to Data: Comparison of Cumulative Distributions



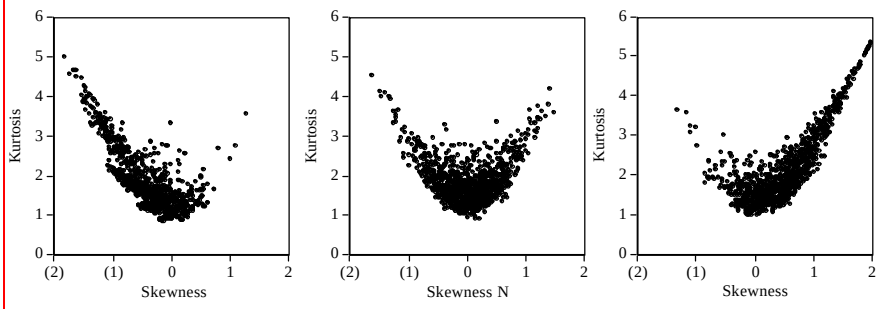
NC STATE UNIVERSITY

Which Distribution is the Correct One?

- Selection of a distribution is subjective
- With small sample sizes, statistical goodness-of-fit techniques have little statistical power
- There is judgment in the selection of:
  - parametric distribution
  - parameter estimation method
  - goodness-of-fit methods
  - specific criteria to use for rejection in a given goodness-of-fit method

NC STATE UNIVERSITY

Bootstrap Simulation of Skewness and Kurtosis to Aid in Selecting a Distribution

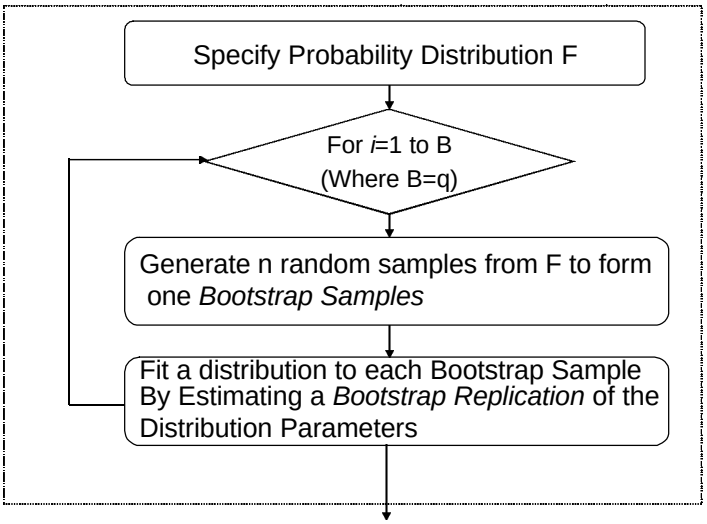


Resampling ***Normal*** ***Lognormal***

Normal Distribution May be a Better Fit
Data Could Be a Sample from a Lognormal Distribution

NC STATE UNIVERSITY

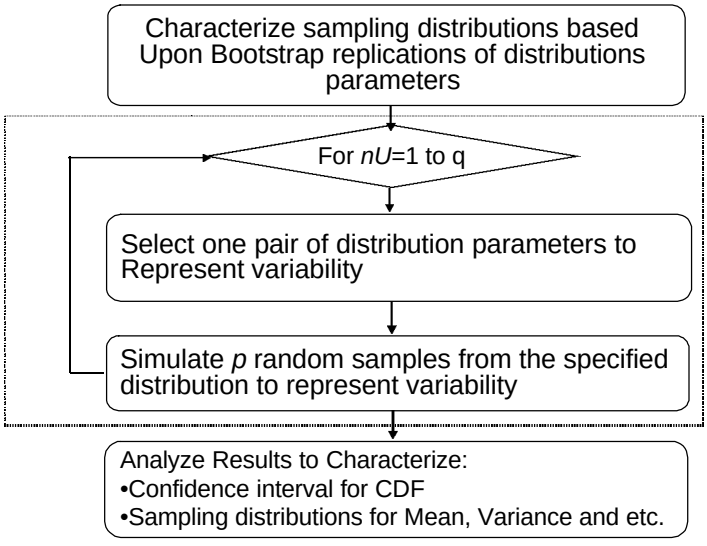
Two-Dimensional Monte Carlo Simulation



Bootstrap Simulation

NC STATE UNIVERSITY

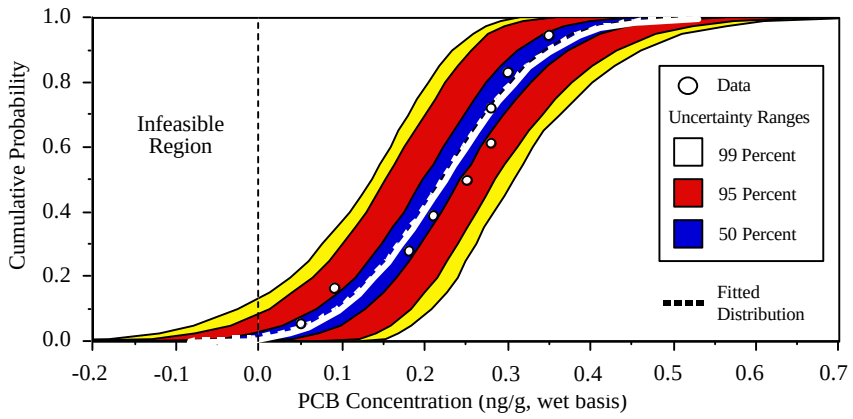
Two-Dimensional Monte Carlo Simulation



Two-Dimensional Simulation of Variability and Uncertainty

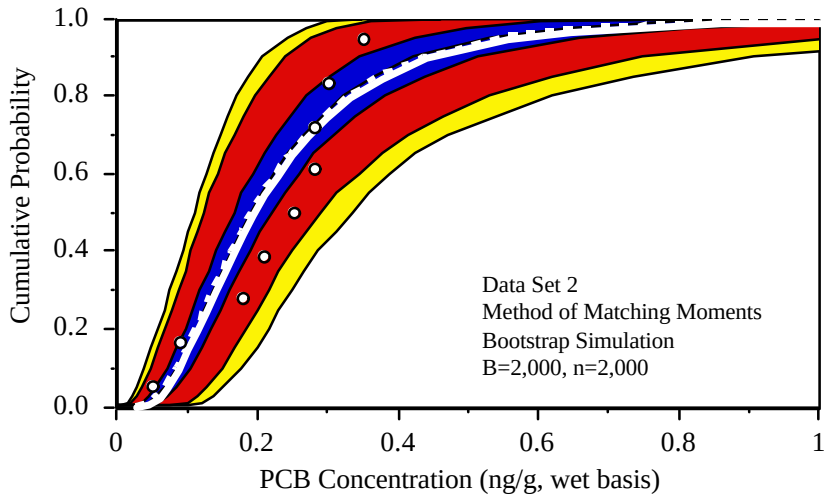
NC STATE UNIVERSITY

Bootstrap Simulation of Confidence Intervals on Fitted Normal Distribution



NC STATE UNIVERSITY

Bootstrap Simulation of Confidence Intervals on Fitted Lognormal Distribution



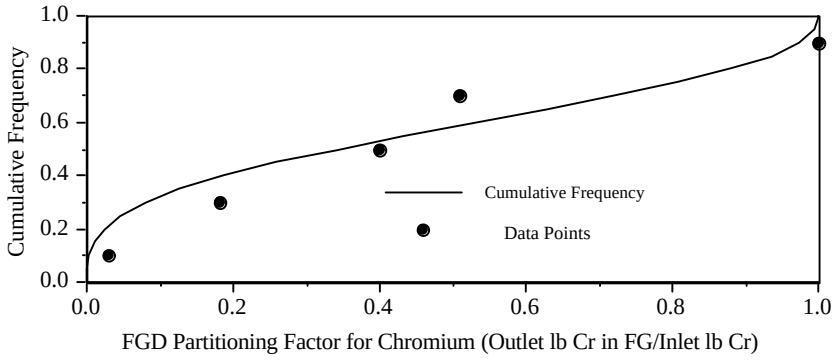
NC STATE UNIVERSITY

Examples of Confidence Intervals: Sensitivity to Selected Distribution

Statistic	Normal Distribution	Lognormal Distribution (MoMM)	Lognormal Distribution (MLE)
5th Percentile of Variability	(-0.03, 0.16)	(0.04, 0.13)	(0.04, 0.13)
50th Percentile of Variability	(0.15, 0.29)	(0.12, 0.29)	(0.13, 0.28)
95th Percentile of Variability	(0.28, 0.48)	(0.28, 1.02)	(0.27, 0.98)
Arithmetic Mean	(0.15, 0.28)	(0.15, 0.37)	(0.15, 0.35)
Arithmetic Variance	(0.0027, 0.020)	(0.0033, 0.10)	(0.0028, 0.09

NC STATE UNIVERSITY

Example 2: Data Set for a Partitioning Factor



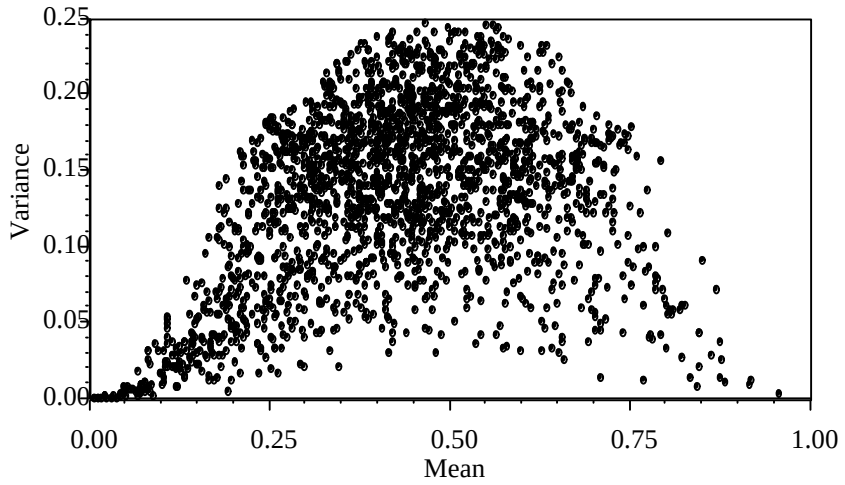
NC STATE UNIVERSITY

Example Case Studies

- Original data set
  - Has a value of 1.0
  - Denoted as **Data Set 2a (DS2a)**
- Alternative Data Set
  - Largest value adjusted from 1.0 to 0.96
  - Denoted as **Data Set 2b (DS2b)**
- Evaluate sensitivity of fitted distributions to this change and to parameter estimation method

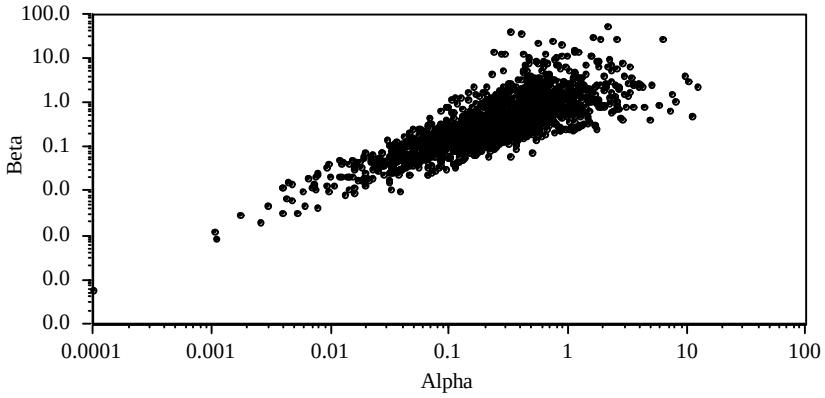
NC STATE UNIVERSITY

Dependence Between Arithmetic Mean and Standard Deviation



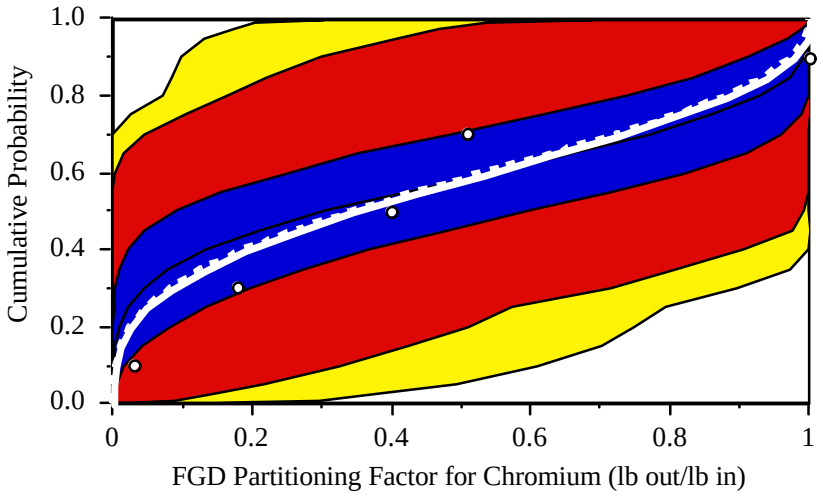
NC STATE UNIVERSITY

Dependence Between Distribution Parameters Alpha and Beta



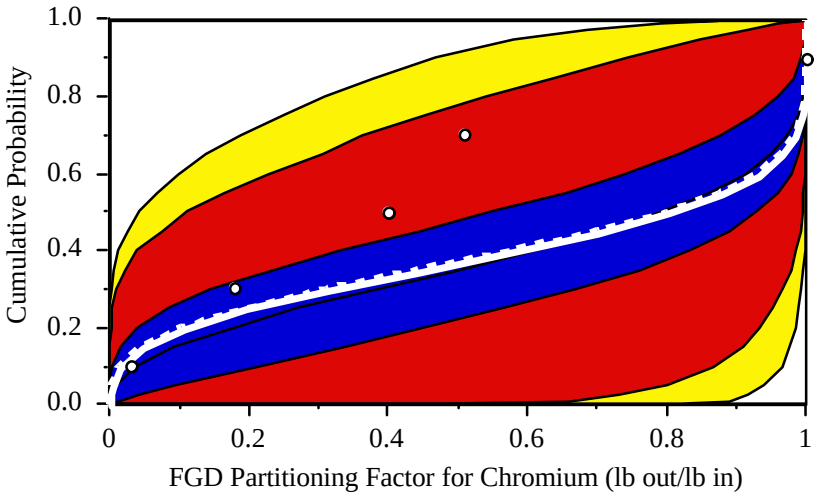
NC STATE UNIVERSITY

Second Order Random Variable for Data Set 2a: Method of Moments



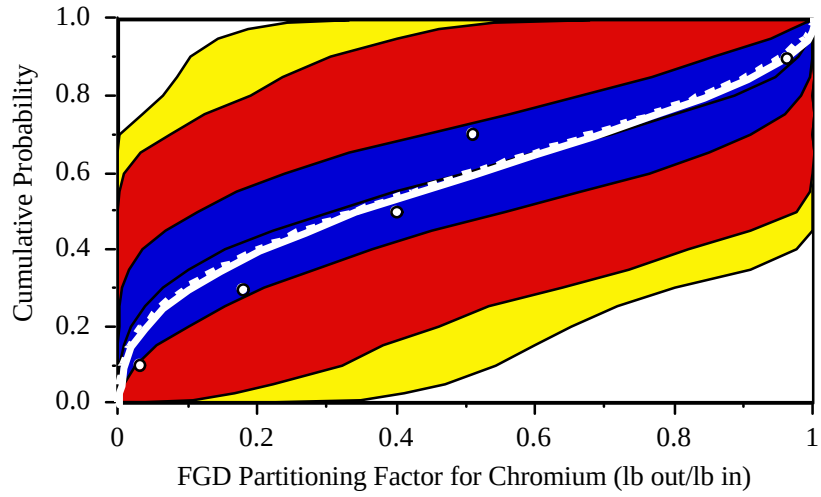
NC STATE UNIVERSITY

Second Order Random Variable for Data Set 2a: Based on MLE



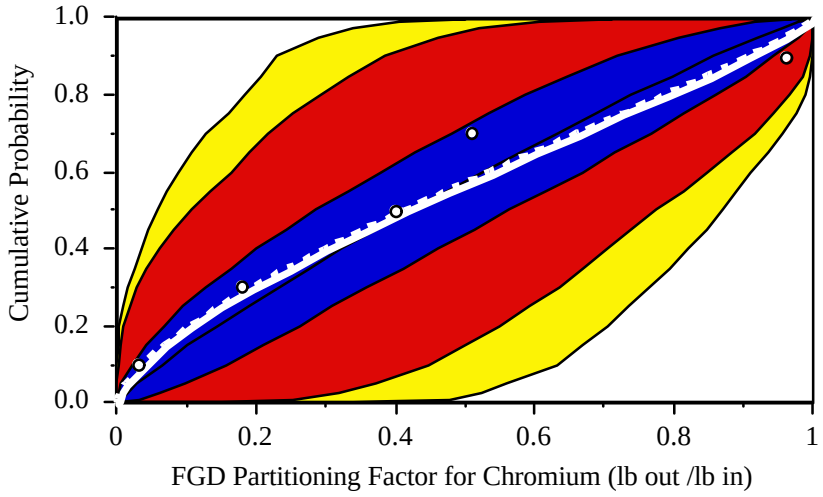
NC STATE UNIVERSITY

Second Order Random Variable for Data Set 2b: Based on MoMM



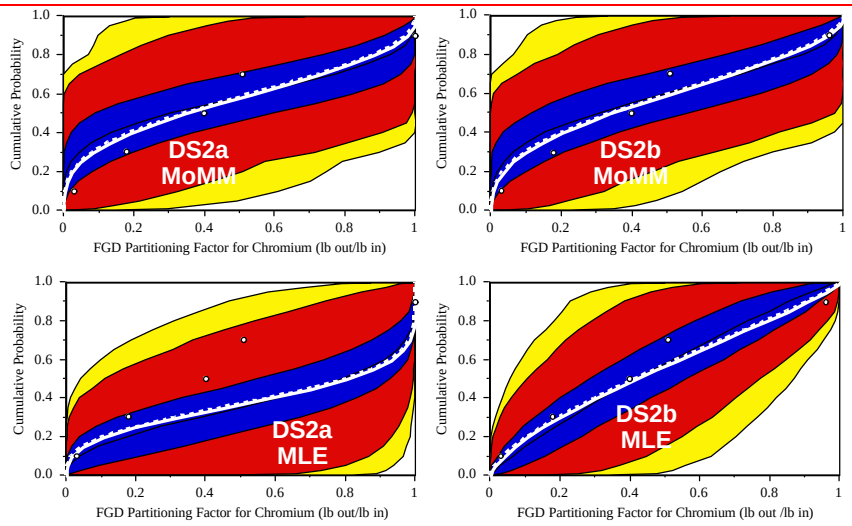
NC STATE UNIVERSITY

Second Order Random Variable for Data Set 2b: Based on MLE



NC STATE UNIVERSITY

Comparison of Four Cases for Data Sets 2a and 2b



NC STATE UNIVERSITY

Data Sets 2a and 2b: Confidence Intervals for Selected Statistics

Statistic	DS3a MoMM	DS3a MLE	DS3b MoMM	DS3b MLE
5th Percentile of Variability	(0, .22)	(0, .81)	(0, .23)	(0, .22)
50th Percentile of Variability	(0, .99)	(0.11, 0.996)	(0, 0.98)	(0.11, 0.78)
95th Percentile of Variability	(0.41, 1.00)	(0.85, 1.00)	(0.40, 1.00)	(0.46, 0.999)
Parameter α	(0.02, 2.22)	(0.16, 7.15)	(0.03, 2.34)	(0.32, 5.79)
Parameter β	(0.02, 5.54)	(0.11, 1.32)	(0.04, 6.14)	(0.37, 7.75)
Arithmetic Mean	(0.10, 0.75)	(0.26, 0.95)	(0.11, 0.73)	(0.18, 0.72)
Arithmetic Variance	(0.013, 0.23)	(0.004, 0.27)	(0.013, 0.23)	(0.015, 0.20)

NC STATE UNIVERSITY

Goodness-of-Fit Tests

- Null hypothesis: data were obtained from the hypothesized distribution
- Require a minimum amount of data (varies for different tests)
- A test statistic is calculated based on the data
- The value of the test statistic is compared to a critical value
- If the test statistic exceeds the critical value, then the null hypothesis is rejected
- One cannot “prove” that a hypothesized distribution is “correct”

NC STATE UNIVERSITY

Chi-Squared Test

- select a hypothesized distribution
- estimate the parameters of the distribution from the data set (need at least 25)
- group the values into cells (or bins) in which each cell has at least five data points
- calculate the probability of obtaining values within the range of each cell based upon the hypothesized distribution
- calculate the expected number of data points that should be in each cell if the hypothesized distribution is acceptable
- calculate a test statistic; and
- evaluate the test statistic

NC STATE UNIVERSITY

Example of Chi-Squared Test: Normal Distribution Fitted to a Data Set (n=25)

Cell Number	End Points of Each Cell		Number of Values in Cell, M_i	Cell Probability, p_i, based on Normal Dist.	Expected Number of Values in Cell, E_i	Test Statistic
	Lower Bound	Upper Bound				
1	0.04	0.09	5	0.0941	2.35	2.98
2	0.09	0.12	5	0.0732	1.83	5.50
3	0.12	0.17	5	0.1432	3.58	0.56
4	0.17	0.27	5	0.2955	7.39	0.77
5	0.27	0.51	5	0.2699	6.75	0.45
Sum of Values:			25	0.876	21.90	10.27

- Test Statistic: $X^2 = \sum_{i=1}^{k} \frac{(M_i - E_i)^2}{E_i}$
- Compare to chi-square distribution with k-r-1 degrees of freedom, k = 5 bins, r = 2 parameters; dof = 2
- Critical value = 6.0, test value = 10.3, reject hypothesis

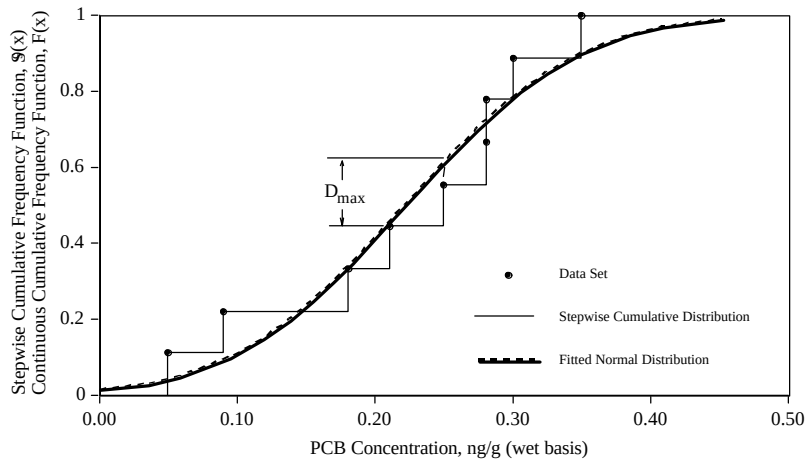
NC STATE UNIVERSITY

Kolmogorov-Smirnov Test

- Comparison between a stepwise empirical CDF and the CDF of the hypothesized distribution.
- Maximum discrepancy in the estimated cumulative probabilities for the two CDFs is identified.
- Maximum discrepancy is then compared to a critical value of the test statistic.
- If the maximum discrepancy is larger than the critical value, then the hypothesized distribution is rejected
- More sensitive near the center of the distribution than at the tails
- Need at least 5 data points

NC STATE UNIVERSITY

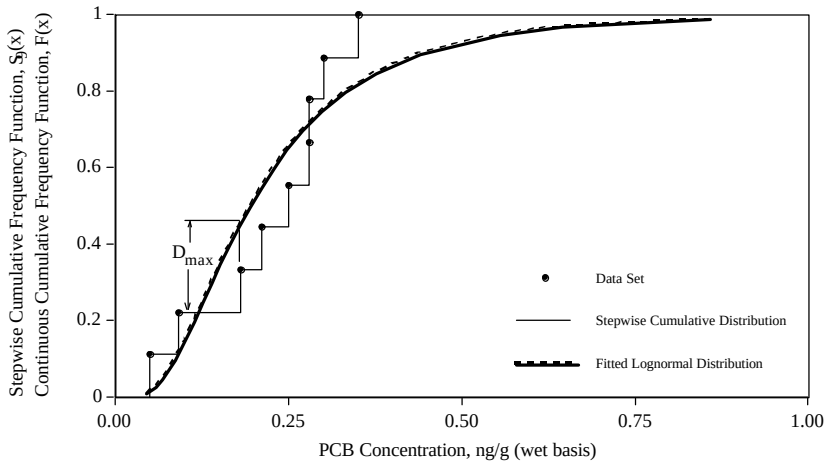
Kolmogorov-Smirnov Test Applied to Leafy Produce Example, Normal Dist.



Maximum difference is 0.17, which is less than critical value of 0.4

NC STATE UNIVERSITY

Kolmogorov-Smirnov Test Applied to Leafy Produce Example, Lognormal



Maximum difference is 0.24, which is less than critical value of 0.4

NC STATE UNIVERSITY

Kolmogorov-Smirnov Test Applied to Leafy Produce Example

- Cannot reject either the Normal or Lognormal as a fit to the data
- Goodness-of-fit tests may lead to inconclusive results

NC STATE UNIVERSITY

Anderson-Darling Test

- A modification of the K-S test
- Gives more weight to the tails than does the K-S test
- The A-D test is not a distribution-free test. For different distributions, A-D test statistics and the corresponding critical values are different

NC STATE UNIVERSITY

Selection of Probabilistic Distribution Models

- Consideration of processes that generate random variable
- Goodness of fit
- The purpose of application of distributions
- Goodness-of-fit tests may lead to inconclusive results
- Bootstrap simulation technique

NC STATE UNIVERSITY

Advice from Hahn and Shapiro (1967)

- One might conclude... that a proper procedure for selecting a distribution is to consider a wide variety of possible models, evaluate each by the methods here described, and assume as correct the one that provides the best fit to the data. However, *no* such approach is being suggested. Where possible, the selection of the model should be based on an understanding of the underlying physical properties... The distributional test then provides a useful mechanism for evaluating the adequacy of the physical interpretation. Only as a last resort is the reverse procedure warranted, and then, only with much care, for, although many models might appear appropriate within the range of the data, they might well be in error in the range for which predictions are desired.[pp 260-261].

NC STATE UNIVERSITY

For More Information

- Most of the examples presented here are from Chapter 5 of:

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

Introduction to AuvTool, Installation and Its Use

NC STATE UNIVERSITY

Junyu (Allen) Zheng, Ph.D
H. Christopher Frey, Ph.D

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

NC STATE UNIVERSITY

Acknowledgement and Disclaimer

- Developed at N.C. State University with support from the Office of Research and Development (ORD) of the U.S. Environmental Protection Agency
- AuvTool 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.

NC STATE UNIVERSITY

AuvTool: Objectives

- To develop a software module named AuvTool (Analysis of Variability and Uncertainty Tool) for use with the EPA Stochastic Human Exposure Dose Simulation (SHEDS) modeling framework
- To implement two-dimensional Monte Carlo method for simultaneously quantifying variability and uncertainty through the AuvTool
- To make the module more generally applicable for some other quantitative analysis fields

NC STATE UNIVERSITY

AuvTool System Development: Design Considerations

- Easily accessible to EPA SHEDS model
- Batch analysis to deal with a large amount of data sets
- Generally applicable for other quantitative analysis fields such as emission estimation and risk assessment
- Extensibility and expansion of AuvTool

NC STATE UNIVERSITY

AuvTool Main Features

- An input sheet similar to spreadsheet
- List of Distributions
  - Normal
  - Beta
  - Weibull
  - Symmetric Triangle
  - Lognormal
  - Gamma
  - Uniform
  - Empirical
  - Mixture normal with two components
  - Mixture lognormal with two components
- Parameter Estimation Methods
  - Matching Moment
  - Maximum Likelihood Estimation (MLE)

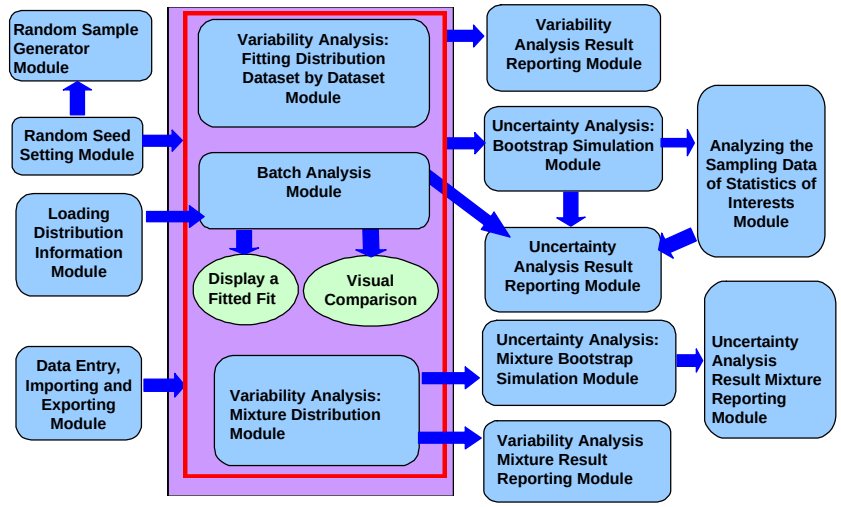
NC STATE UNIVERSITY

AuvTool Main Features (Cont'd)

- Batch analysis
  - To automatically help user choose the best distributions
- Bootstrap simulation and two-dimensional simulation
  - Single component distributions,
  - mixture distributions
- Statistical Goodness of fit tests
  - Kolmogorov-Smirnov test (K-S)
  - Anderson-Darling test (A-D)
- Instant graphical presentation and tabular summarization of results

NC STATE UNIVERSITY

AuvTool Software Implementation: Structure Design



NC STATE UNIVERSITY

AuvTool Software Development Environment and Tools

- Visual C++ 6.0
  - To implement all calculations and graphic user interface
- Graphics Server 5.0A
  - To present calculation results in the visual graphic form
- Spread 3.0
  - To provide a spreadsheet for user data input and calculation result output in tabular form
- Based on Windows 98 development environment
- An object-oriented programming technology is used
  - More than 45 C++ classes are designed

NC STATE UNIVERSITY

AuvTool Installation

1. Place the CD-ROM in your CD-ROM drive;
2. Click the Start button;
3. Choose Run… from the Start menu;
4. Type "X:\ XXX\" SETUP.EXE" where "X:\ " is the drive and directory to which you copied the installation files. The Installation Program will begin. Follow the instructions on the screen.

You also can install AuvTool as follows:

1. Place the AuvTool CD-ROM in the CD-ROM drive;
2. Double-click the My Computer icon on the desktop;
3. Double-click the CD-ROM drive in the My Computer window; and
4. Double-click the "SETUP.EXE" on the CD-ROM.

The Installation program will start. Follow the instructions on the screen.

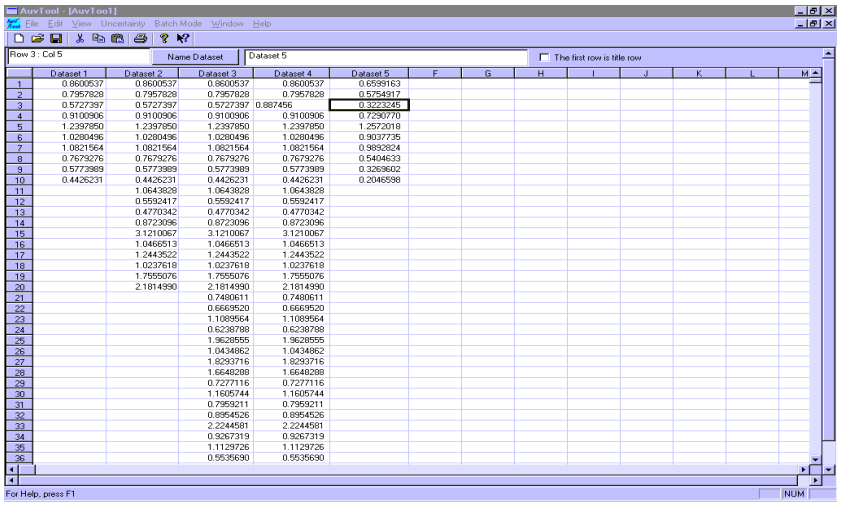
NC STATE UNIVERSITY

The Use of AuvTool

- Online help system in the AuvTool
- A PDF file of AuvTool's user guide is available in the accompanying CD disk
- An demo example

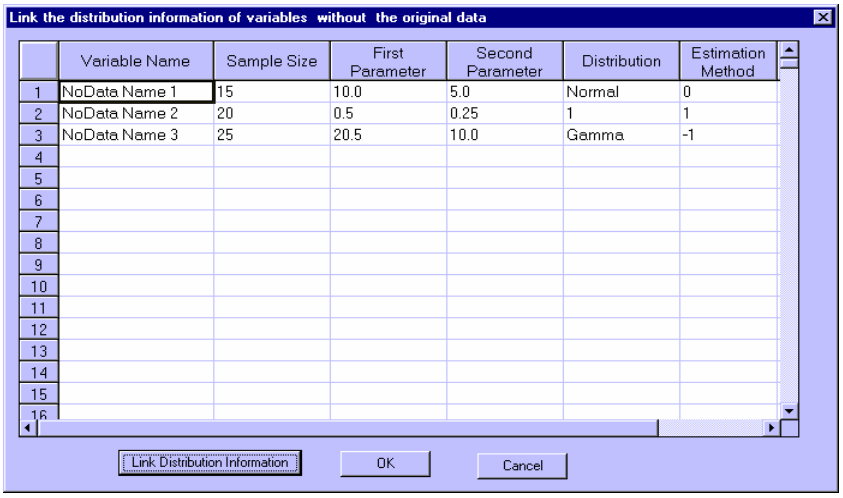
NC STATE UNIVERSITY

Data Entry, Importing and Exporting



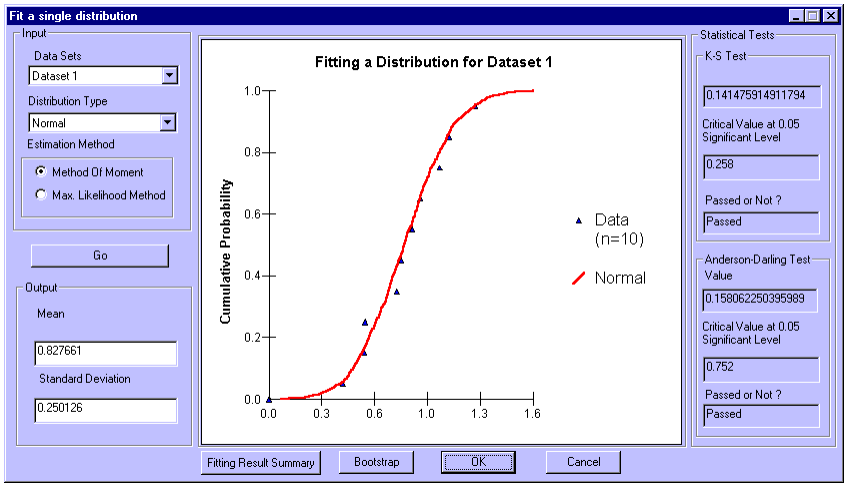
NC STATE UNIVERSITY

Loading Known Distribution Information



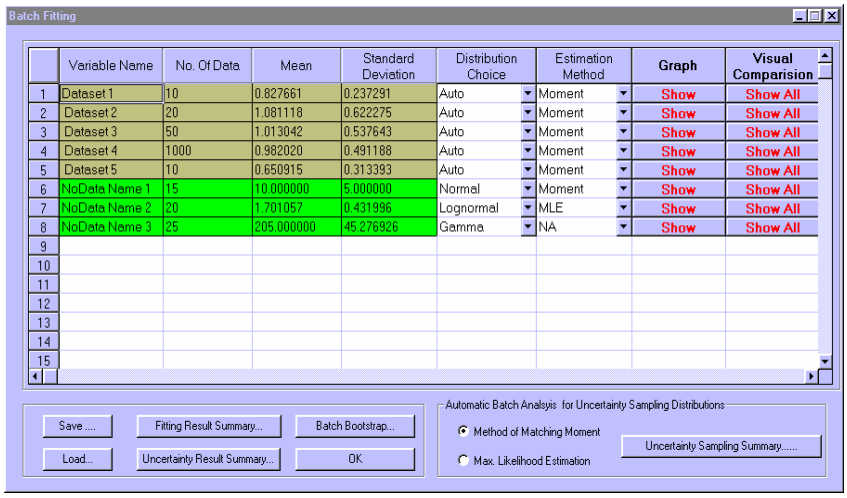
NC STATE UNIVERSITY

Fitting a Distribution for Variability: Dataset by Dataset



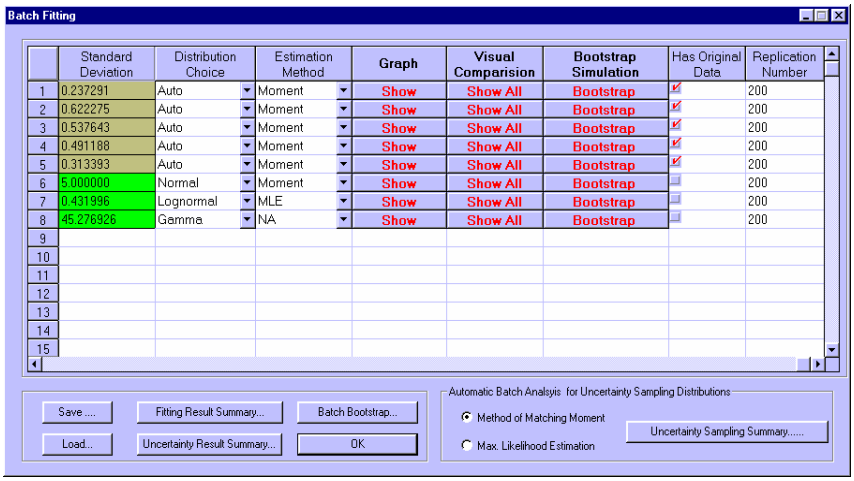
NC STATE UNIVERSITY

Batch Analysis (1)



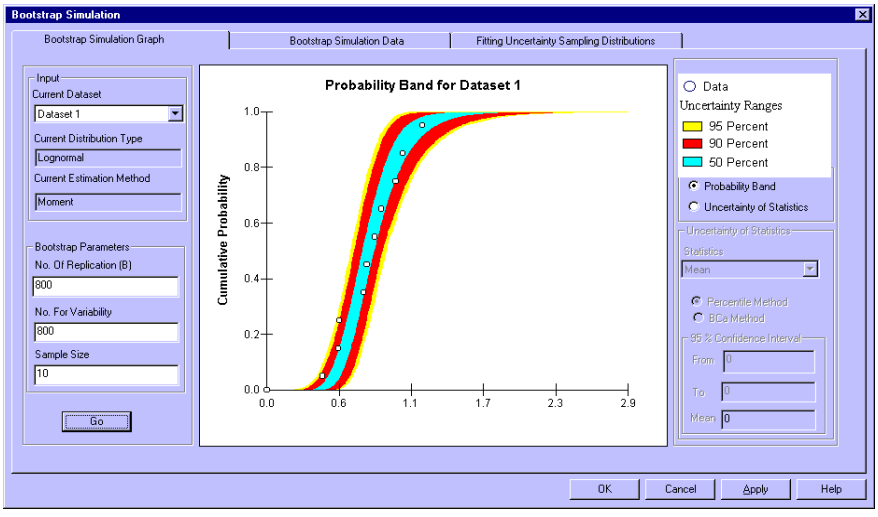
NC STATE UNIVERSITY

Batch Analysis (2)



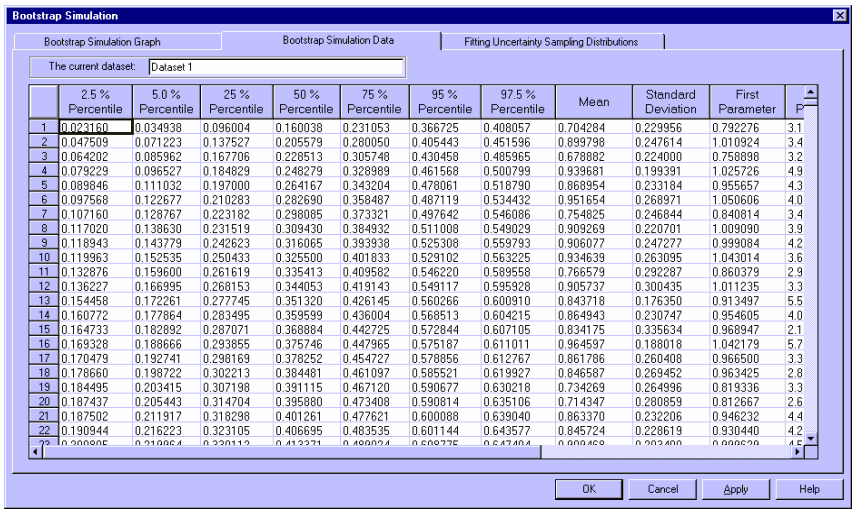
NC STATE UNIVERSITY

Bootstrap Simulation: Probability Band Graph



NC STATE UNIVERSITY

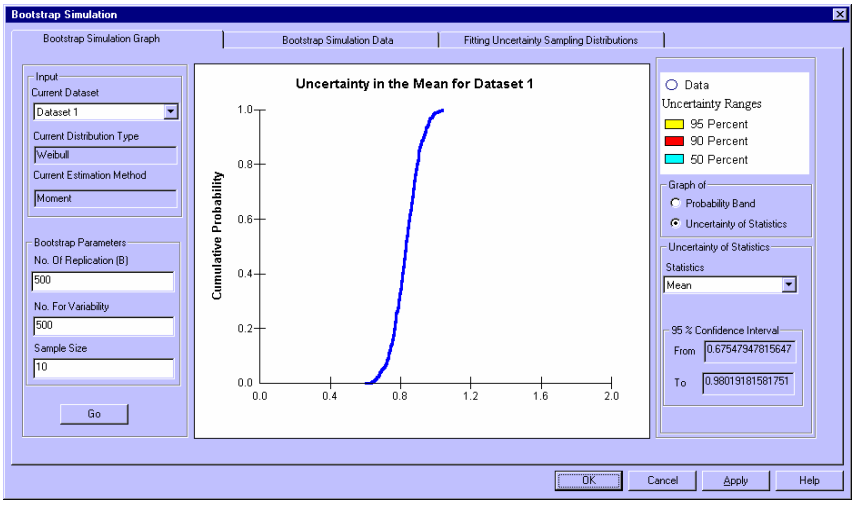
Bootstrap Simulation- Data



	2.5 % Percentile	5.0 % Percentile	25 % Percentile	50 % Percentile	75 % Percentile	95 % Percentile	97.5 % Percentile	Mean	Standard Deviation	First Parameter
1	0.023160	0.034938	0.096004	0.160038	0.231053	0.366725	0.408057	0.704284	0.229956	0.792276
2	0.047509	0.071223	0.137527	0.205579	0.280050	0.405443	0.451596	0.899798	0.247614	1.010924
3	0.064202	0.085962	0.167706	0.228513	0.305748	0.430458	0.485965	0.678882	0.224000	0.758898
4	0.079229	0.096527	0.184829	0.248279	0.328989	0.461568	0.500799	0.939681	0.199391	1.025726
5	0.089846	0.111032	0.197000	0.264167	0.343204	0.478061	0.518790	0.868954	0.233184	0.955657
6	0.097568	0.122677	0.210283	0.282690	0.358487	0.487119	0.534432	0.951654	0.268971	1.050606
7	0.107160	0.128767	0.223182	0.298085	0.373321	0.497642	0.546086	0.754825	0.246844	0.840814
8	0.117020	0.138630	0.231519	0.309430	0.384932	0.511008	0.549029	0.909269	0.220701	1.009090
9	0.118943	0.143779	0.242623	0.316065	0.393938	0.525308	0.559793	0.906077	0.247277	0.999084
10	0.119963	0.152535	0.250433	0.325500	0.401833	0.529102	0.563225	0.934639	0.263095	1.043014
11	0.132876	0.159600	0.261619	0.335413	0.409582	0.546220	0.589558	0.766579	0.292287	0.860379
12	0.136227	0.166995	0.268153	0.344053	0.419143	0.549117	0.595928	0.905737	0.300435	1.011235
13	0.154458	0.172261	0.277745	0.351320	0.426145	0.560266	0.600910	0.843718	0.176350	0.913497
14	0.160772	0.177864	0.283495	0.359599	0.436004	0.568513	0.604215	0.864943	0.230747	0.954605
15	0.164733	0.182892	0.287071	0.368884	0.442725	0.572844	0.607105	0.834175	0.335634	0.968947
16	0.169328	0.188666	0.293855	0.375746	0.447965	0.575187	0.611011	0.964597	0.188018	1.042179
17	0.170479	0.192741	0.298169	0.378252	0.454727	0.578856	0.612767	0.861786	0.260408	0.966500
18	0.178660	0.198722	0.302213	0.384481	0.461097	0.585521	0.619927	0.846587	0.269452	0.963425
19	0.184495	0.203415	0.307198	0.391115	0.467120	0.590677	0.630218	0.734269	0.264996	0.819336
20	0.187437	0.205443	0.314704	0.395880	0.473408	0.590814	0.635106	0.714347	0.280859	0.812667
21	0.187502	0.211917	0.318298	0.401261	0.477621	0.600088	0.639040	0.863370	0.232206	0.946232
22	0.190944	0.216223	0.323105	0.406695	0.483535	0.601144	0.643577	0.845724	0.228619	0.930440

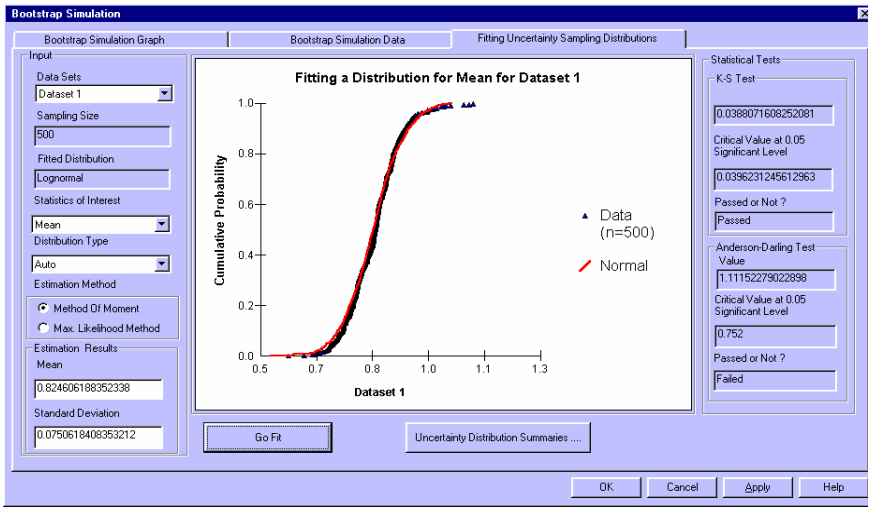
NC STATE UNIVERSITY

Bootstrap: Uncertainty Graph



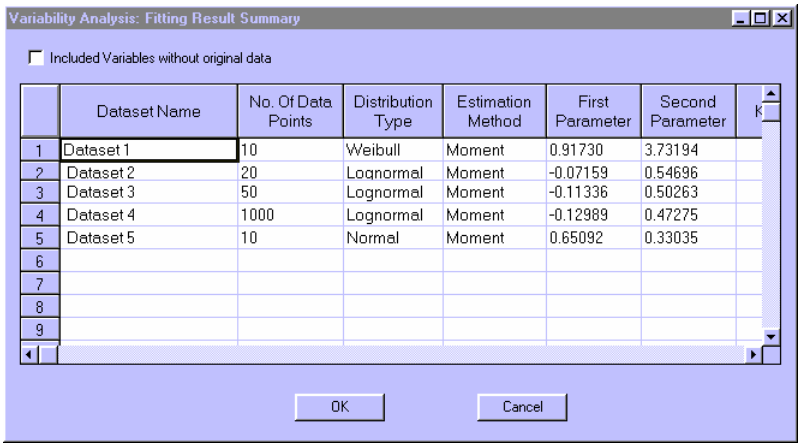
NC STATE UNIVERSITY

Analyzing the Sampling Data of Statistics of Interests



NC STATE UNIVERSITY

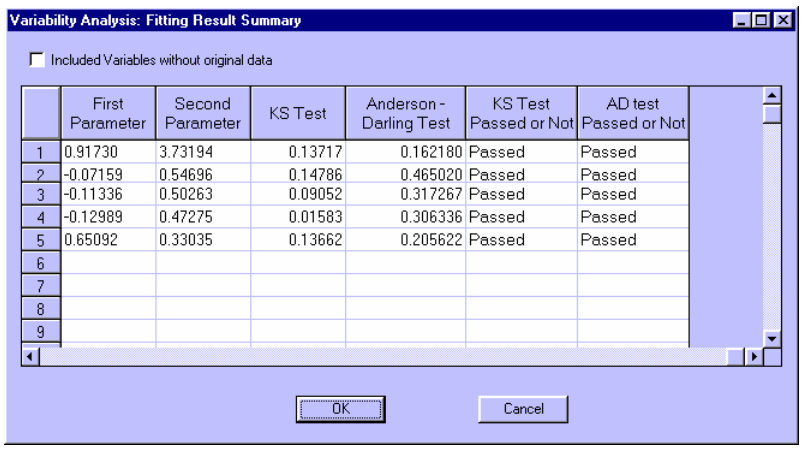
Variability Analysis Result Reporting (1)



	Dataset Name	No. Of Data Points	Distribution Type	Estimation Method	First Parameter	Second Parameter
1	Dataset 1	10	Weibull	Moment	0.91730	3.73194
2	Dataset 2	20	Lognormal	Moment	-0.07159	0.54696
3	Dataset 3	50	Lognormal	Moment	-0.11336	0.50263
4	Dataset 4	1000	Lognormal	Moment	-0.12989	0.47275
5	Dataset 5	10	Normal	Moment	0.65092	0.33035
6						
7						
8						
9						

NC STATE UNIVERSITY

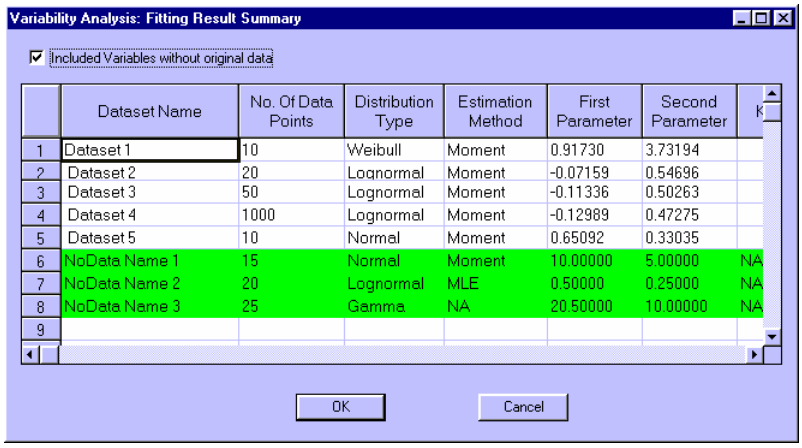
Variability Analysis Result Reporting (2)



	First Parameter	Second Parameter	KS Test	Anderson - Darling Test	KS Test Passed or Not	AD test Passed or Not
1	0.91730	3.73194	0.13717	0.162180	Passed	Passed
2	-0.07159	0.54696	0.14786	0.465020	Passed	Passed
3	-0.11336	0.50263	0.09052	0.317267	Passed	Passed
4	-0.12989	0.47275	0.01583	0.306336	Passed	Passed
5	0.65092	0.33035	0.13662	0.205622	Passed	Passed
6						
7						
8						
9						

NC STATE UNIVERSITY

Variability Analysis Result Reporting (3)



	Dataset Name	No. Of Data Points	Distribution Type	Estimation Method	First Parameter	Second Parameter	K
1	Dataset 1	10	Weibull	Moment	0.91730	3.73194	
2	Dataset 2	20	Lognormal	Moment	-0.07159	0.54696	
3	Dataset 3	50	Lognormal	Moment	-0.11336	0.50263	
4	Dataset 4	1000	Lognormal	Moment	-0.12989	0.47275	
5	Dataset 5	10	Normal	Moment	0.65092	0.33035	
6	NoData Name 1	15	Normal	Moment	10.00000	5.00000	NA
7	NoData Name 2	20	Lognormal	MLE	0.50000	0.25000	NA
8	NoData Name 3	25	Gamma	NA	20.50000	10.00000	NA
9							

NC STATE UNIVERSITY

Uncertainty Analysis Result Reporting (1)

Uncertainty Analysis Summary

	Dataset Name	No. Of Data	Distribution Type	Estimation Method	Mean 2.5% Percentile	Mean Mean	Mean 97.5 Percentile	Std. 2.5%
1	Dataset 1	10	Weibull	Moment	0.688012	0.826736	0.977089	0.129
2	Dataset 2	20	Lognormal	Moment	0.835350	1.078356	1.385800	0.337
3	Dataset 3	50	Lognormal	Moment	0.883052	1.009503	1.147866	0.376
4	Dataset 4	1000	Lognormal	Moment	0.951584	0.982140	1.013923	0.448
5	Dataset 5	10	Normal	Moment	0.415494	0.651159	0.868226	0.172
6	NoData Name 1	15	Normal	Moment	7.497081	9.962685	12.793174	3.015
7	NoData Name 2	20	Lognormal	Moment	1.519883	1.703602	1.893432	0.280
8	NoData Name 3	25	Gamma	Moment	187.924497	205.114160	221.796368	31.04
9								
10								
11								
12								
13								
14								
15								
16								
17								

OK Cancel

NC STATE UNIVERSITY

Uncertainty Analysis Result Reporting (2)

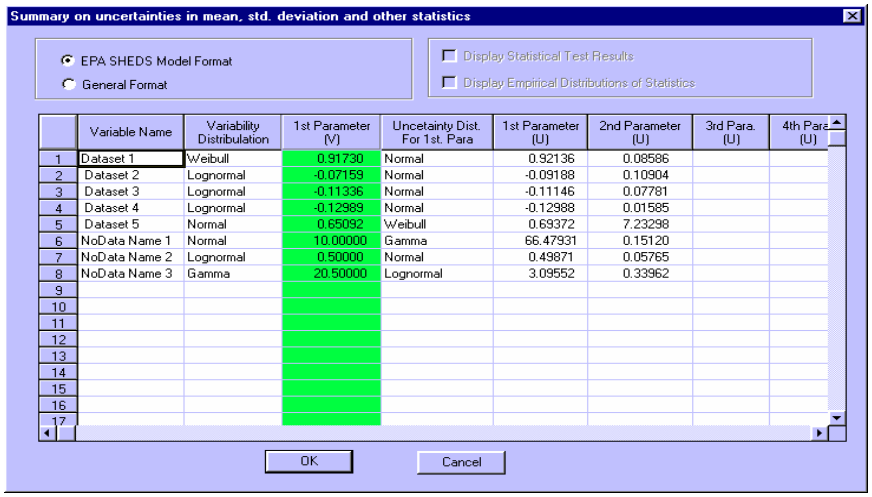
Uncertainty Analysis Summary

	Estimation Method	Mean 2.5% Percentile	Mean Mean	Mean 97.5 Percentile	Std. Deviation 2.5% Percentile	Std. Deviation Mean	Std. Deviation 97.5% Percentile
1	Moment	0.688012	0.826736	0.977089	0.129039	0.237607	0.347369
2	Moment	0.835350	1.078356	1.385800	0.337081	0.606995	1.051127
3	Moment	0.883052	1.009503	1.147866	0.376882	0.522093	0.733852
4	Moment	0.951584	0.982140	1.013923	0.448970	0.490640	0.536754
5	Moment	0.415494	0.651159	0.868226	0.172858	0.320810	0.469982
6	Moment	7.497081	9.962685	12.793174	3.015366	4.914352	6.864976
7	Moment	1.519883	1.703602	1.893432	0.280687	0.424432	0.628553
8	Moment	187.924497	205.114160	221.796368	31.040363	44.766731	59.796766
9							
10							
11							
12							
13							
14							
15							
16							
17							

OK Cancel

NC STATE UNIVERSITY

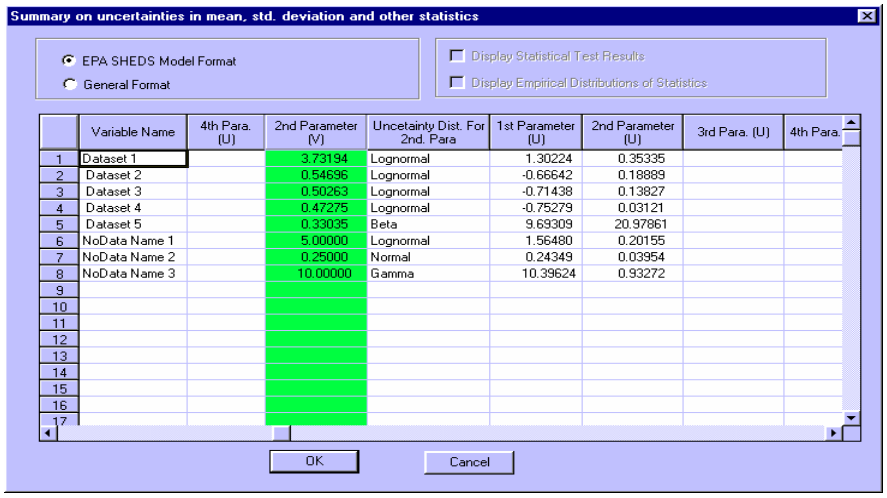
Uncertainty Analysis Result Reporting (3): SHEDS Model Format



	Variable Name	Variability Distribulation	1st Parameter (V)	Uncetainty Dist. For 1st. Para	1st Parameter (U)	2nd Parameter (U)	3rd Para. (U)	4th Para. (U)
1	Dataset 1	Weibull	0.91730	Normal	0.92136	0.08586		
2	Dataset 2	Lognormal	-0.07159	Normal	-0.09188	0.10904		
3	Dataset 3	Lognormal	-0.11336	Normal	-0.11146	0.07781		
4	Dataset 4	Lognormal	-0.12989	Normal	-0.12988	0.01585		
5	Dataset 5	Normal	0.65092	Weibull	0.69372	7.23298		
6	NoData Name 1	Normal	10.00000	Gamma	66.47931	0.15120		
7	NoData Name 2	Lognormal	0.50000	Normal	0.49871	0.05765		
8	NoData Name 3	Gamma	20.50000	Lognormal	3.09552	0.33962		
9								
10								
11								
12								
13								
14								
15								
16								
17								

NC STATE UNIVERSITY

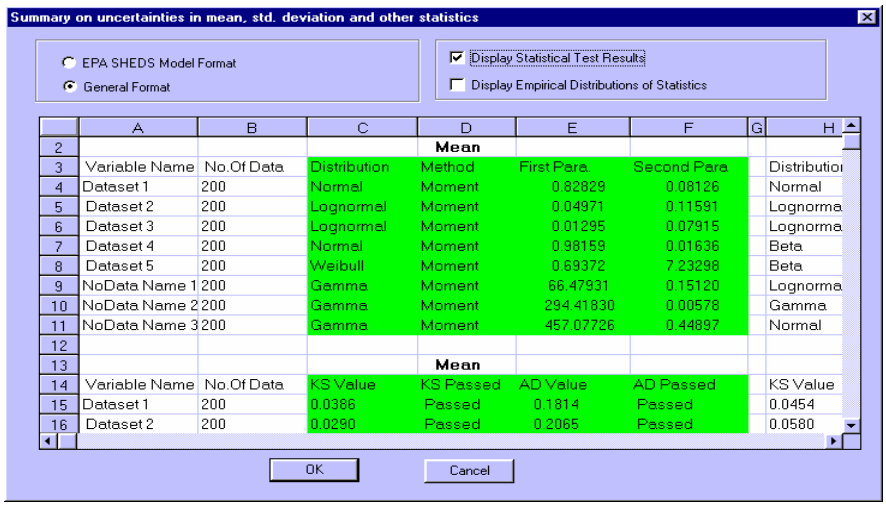
Uncertainty Analysis Result Reporting (4): SHEDS Model Format



	Variable Name	4th Para. (U)	2nd Parameter (V)	Uncetainty Dist. For 2nd. Para	1st Parameter (U)	2nd Parameter (U)	3rd Para. (U)	4th Para.
1	Dataset 1		3.73194	Lognormal	1.30224	0.35335		
2	Dataset 2		0.54696	Lognormal	-0.66642	0.18889		
3	Dataset 3		0.50263	Lognormal	-0.71438	0.13827		
4	Dataset 4		0.47275	Lognormal	-0.75279	0.03121		
5	Dataset 5		0.33035	Beta	9.69309	20.97861		
6	NoData Name 1		5.00000	Lognormal	1.56480	0.20155		
7	NoData Name 2		0.25000	Normal	0.24349	0.03954		
8	NoData Name 3		10.00000	Gamma	10.39624	0.93272		
9								
10								
11								
12								
13								
14								
15								
16								
17								

NC STATE UNIVERSITY

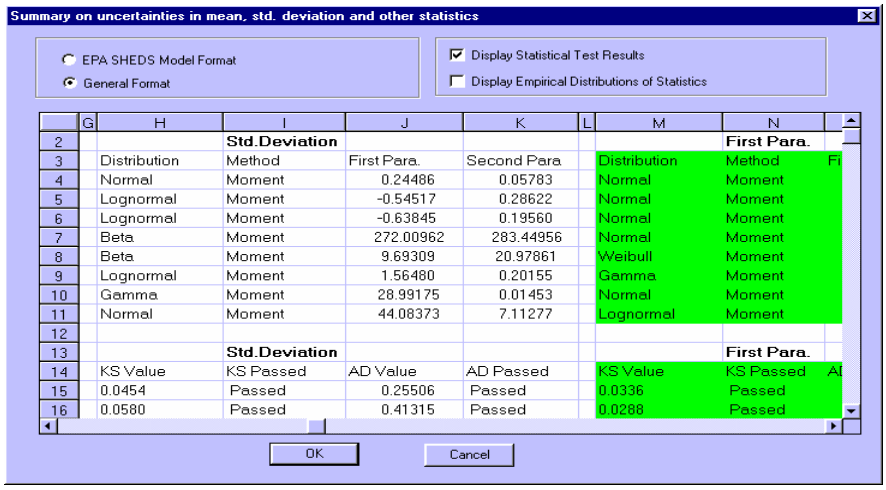
Uncertainty Analysis Result Reporting (5): General Format



	A	B	C	D	E	F	G	H
2				**Mean**				
3	Variable Name	No.Of Data	Distribution	Method	First Para.	Second Para		Distributio
4	Dataset 1	200	Normal	Moment	0.82829	0.08126		Normal
5	Dataset 2	200	Lognormal	Moment	0.04971	0.11591		Lognorma
6	Dataset 3	200	Lognormal	Moment	0.01295	0.07915		Lognorma
7	Dataset 4	200	Normal	Moment	0.98159	0.01636		Beta
8	Dataset 5	200	Weibull	Moment	0.69372	7.23298		Beta
9	NoData Name 1	200	Gamma	Moment	66.47931	0.15120		Lognorma
10	NoData Name 2	200	Gamma	Moment	294.41830	0.00578		Gamma
11	NoData Name 3	200	Gamma	Moment	457.07726	0.44897		Normal
12								
13				**Mean**				
14	Variable Name	No.Of Data	KS Value	KS Passed	AD Value	AD Passed		KS Value
15	Dataset 1	200	0.0386	Passed	0.1814	Passed		0.0454
16	Dataset 2	200	0.0290	Passed	0.2065	Passed		0.0580

NC STATE UNIVERSITY

Uncertainty Analysis Result Reporting (6): General Format



	G	H	I	J	K	L	M	N
2			**Std.Deviation**					**First Para.**
3		Distribution	Method	First Para.	Second Para		Distribution	Method
4		Normal	Moment	0.24486	0.05783		Normal	Moment
5		Lognormal	Moment	-0.54517	0.28622		Normal	Moment
6		Lognormal	Moment	-0.63845	0.19560		Normal	Moment
7		Beta	Moment	272.00962	283.44956		Normal	Moment
8		Beta	Moment	9.69309	20.97861		Weibull	Moment
9		Lognormal	Moment	1.56480	0.20155		Gamma	Moment
10		Gamma	Moment	28.99175	0.01453		Normal	Moment
11		Normal	Moment	44.08373	7.11277		Lognormal	Moment
12								
13			**Std.Deviation**					**First Para.**
14		KS Value	KS Passed	AD Value	AD Passed		KS Value	KS Passed
15		0.0454	Passed	0.25506	Passed		0.0336	Passed
16		0.0580	Passed	0.41315	Passed		0.0288	Passed

NC STATE UNIVERSITY

Conclusion

- Implemented a general tool for quantifying variability and uncertainty in model inputs
- Provided the required variability and uncertainty inputs to the EPA/SHEDS model
- Can be generally used in any application where characterization of variability and uncertainty for datasets is needed.

NC STATE UNIVERSITY

More about AuvTool

- Not a commercial product
- Provided “as is” as a research tool
- Most but not all capabilities described in this workshop are implemented
- No warrantees of any kind
- It has been tested and found to work the best on Windows 98/ME.
- No formal technical support
- No resources at this time for technical support or further modification

NC STATE UNIVERSITY

Potential Bugs in AuvTool

- Program may crash when doing bootstrap simulation (especially for mixture distributions in Windows XP version)
- Importing Excel files or exporting sheets to Excel files may not be successful

NC STATE UNIVERSITY

Quantification of Variability and Uncertainty for Censored Air Toxics Emissions Data Sets

NC STATE UNIVERSITY

H. Christopher Frey, Ph.D
Yuchao Zhao

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

NC STATE UNIVERSITY

Motivations: Censored Datasets

- Toxic air pollutants pose human health risks in urban areas
- Quantification of variability and uncertainty for emissions from air toxics is needed for human exposure and risk analysis
- Emission data sets for urban air toxics often contain several observations as below detection limit, which are referred to as “censored”
  - Single detection limit
  - Multiple detection limits

NC STATE UNIVERSITY

Objectives

- To fit parametric distributions using Maximum Likelihood Estimation (MLE) to censored data sets
- To quantify variability and uncertainty for censored data sets using empirical bootstrap simulations
- To test and apply this method
- To compare with conventional approaches

NC STATE UNIVERSITY

Conventional Approaches for Handling Non-Detects When Calculating the Mean

- Use values only above Detection Limit (DL)
- Replace values below DL with zero
- Replace values below DL by DL/2
- Replace values below DL by DL

NC STATE UNIVERSITY

Alternative to Conventional Approaches

- Fit a parametric distribution to censored data using Maximum Likelihood Estimation (MLE)
- MLE is asymptotically unbiased
- Fitted distribution is the best estimate of variability
- Can estimate mean from the fitted distribution
- Also can estimate other statistics (e.g., standard deviation)
- Can quantify uncertainty because of random sampling error
- Bootstrap simulation can be used to quantify uncertainty

NC STATE UNIVERSITY

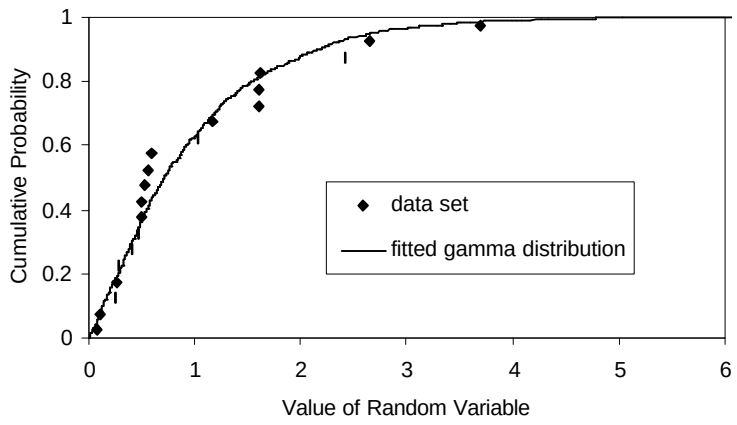
Maximize the Likelihood Function for Parametric Distribution Fitted to Censored Data

$$L(q_1,q_2,...,q_k)=\prod_{i=1}^{n} f(x_i \mid q_1,q_2,..q_k)\left\{\prod_{m=1}^{p}\left(\prod_{j=1}^{ND_m} F(DL_m \mid q_1,q_2,...,q_k)\right)\right\}$$

L	=	Likelihood function
DL_m	=	The m^{th} detection limit
f	=	Probability density function
F	=	Cumulative distribution function
ND_m	=	Number of non-detects corresponding to detection limit DL_m, where, m = 1, 2, …, p
p	=	Number of detection limits
x_i	=	Detected data point, where, i = 1, 2, …, n
$q_1,q_2,...,q_k$	=	Parameters of the distribution

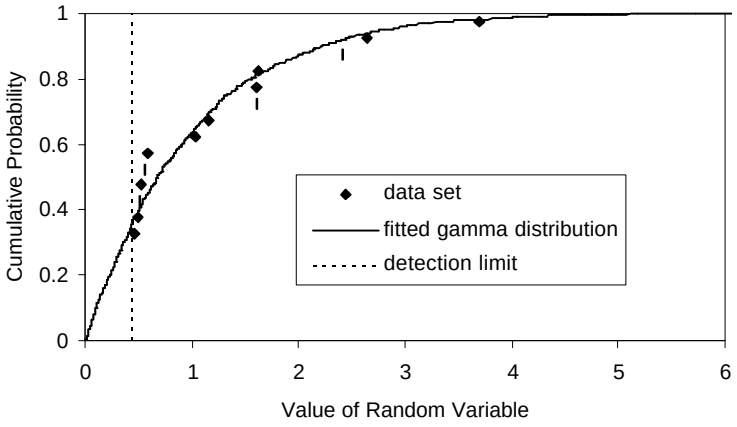
NC STATE UNIVERSITY

Example Results of MLE: Fitted to Gamma, No Censoring



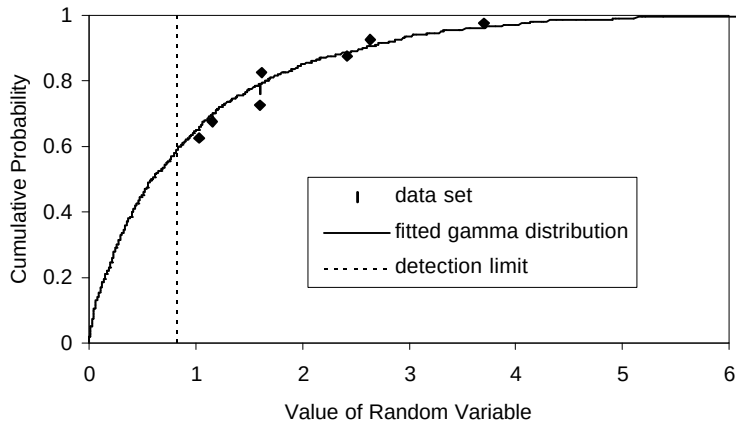
NC STATE UNIVERSITY

Example Results of MLE: Fitted to Gamma, 30% Censoring



NC STATE UNIVERSITY

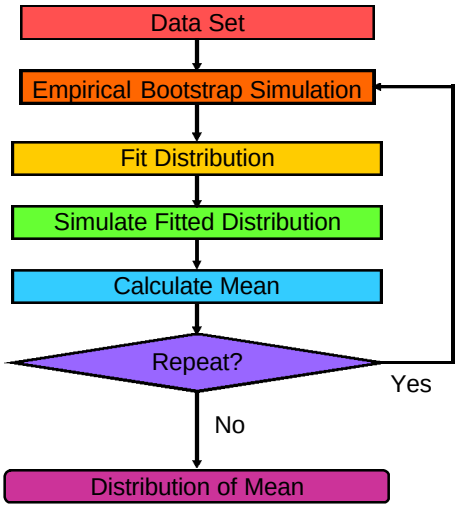
Example Results of MLE: Fitted to Gamma, 60% Censoring



NC STATE UNIVERSITY

Methodology

- Scheme of quantification of uncertainty and variability for censored data sets



NC STATE UNIVERSITY

Methodology: Empirical Bootstrap Simulation

sample at random with replacement

$(X_1, X_2, \ldots, X_n)$ → $(X_1^*, X_2^*, \ldots, X_n^*)$

Original data set

Empirical bootstrap sample

X_i = original data point

X_i^* = random value from original data

i = 1, n

n = sample size of original data

NC STATE UNIVERSITY

Empirical Bootstrap Simulation: Example

Sample at random with replacement

(1, 3, 6, 9) →

Original data set

(3, 1, 9, 1)
(9, 3, 6, 9)
(1, 6, 1, 3)
(9, 1, 6, 9)

Empirical Bootstrap Samples

- Repeat B times, where B = 500
- Each time, calculate a statistic (e.g. distribution parameters)
- The B values of the statistic represent uncertainty

NC STATE UNIVERSITY

Methodology: Empirical Bootstrap Simulation for Censored Data Sets

$$\left[\begin{pmatrix} x_1 \\ d_1 \end{pmatrix}\begin{pmatrix} x_2 \\ d_2 \end{pmatrix}\cdots\begin{pmatrix} x_n \\ d_n \end{pmatrix}\right] \xrightarrow{\text{Sample with replacement}} \left[\begin{pmatrix} x_1^* \\ d_1^* \end{pmatrix}\begin{pmatrix} x_2^* \\ d_2^* \end{pmatrix}\cdots\begin{pmatrix} x_n^* \\ d_n^* \end{pmatrix}\right]$$

- In the original data set, either a true value is given for detected points or the detection limit is given for censored point
- An Indicator symbol d_i (1 or 0) is used to indicate the status of x_i
- Randomly sample both x_i and d_i at the same time

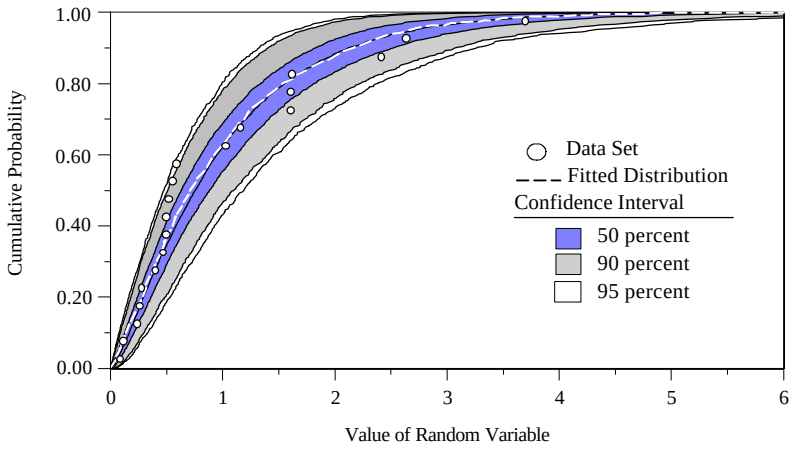
NC STATE UNIVERSITY

Test Case 1

- 20 synthetic data points were generated from a gamma distribution (mean = 1 and standard deviation = 1)
- 0%, 30% and 60% censoring
- Gamma and lognormal distributions were fit to the censored empirical bootstrap samples

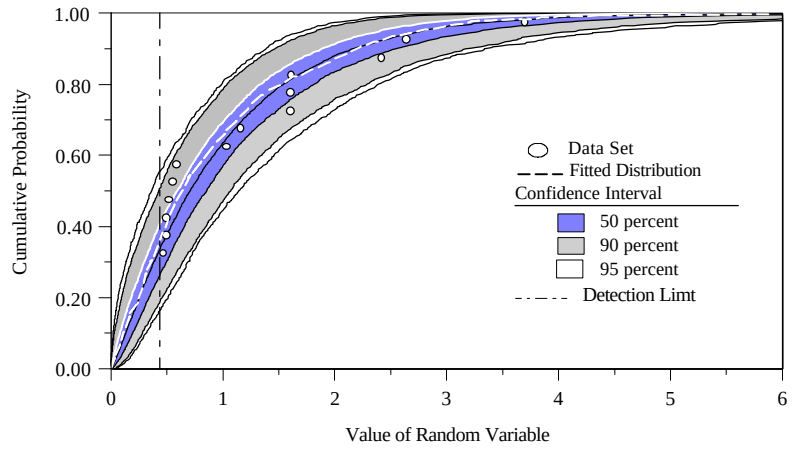
NC STATE UNIVERSITY

Results: Fitted Gamma Distribution, No Censoring



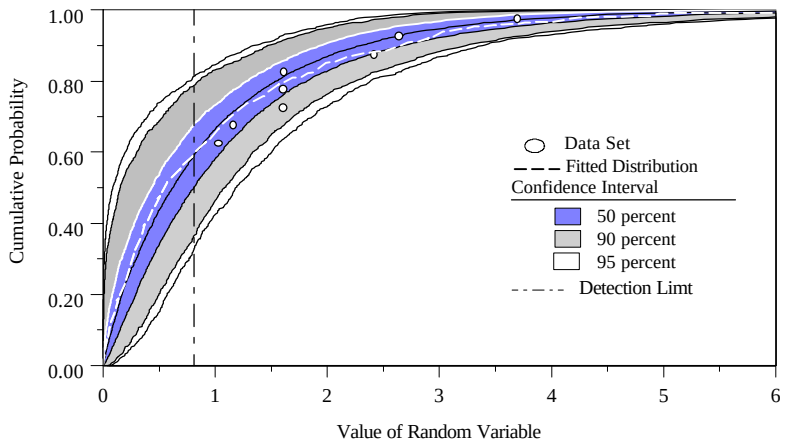
NC STATE UNIVERSITY

Results: Fitted Gamma Distribution, 30% Censoring



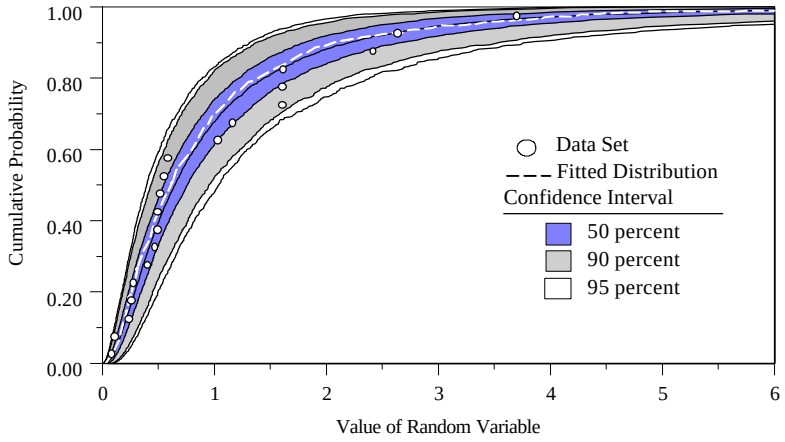
NC STATE UNIVERSITY

Results: Fitted Gamma Distribution, 60% Censoring



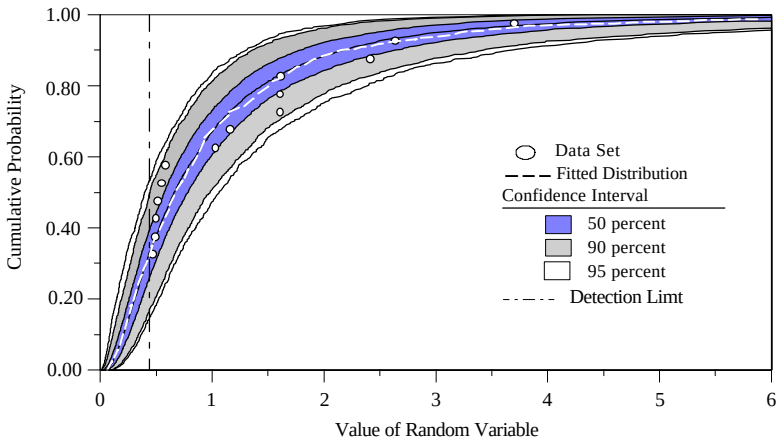
NC STATE UNIVERSITY

Results: Fitted Lognormal Distribution, No Censoring



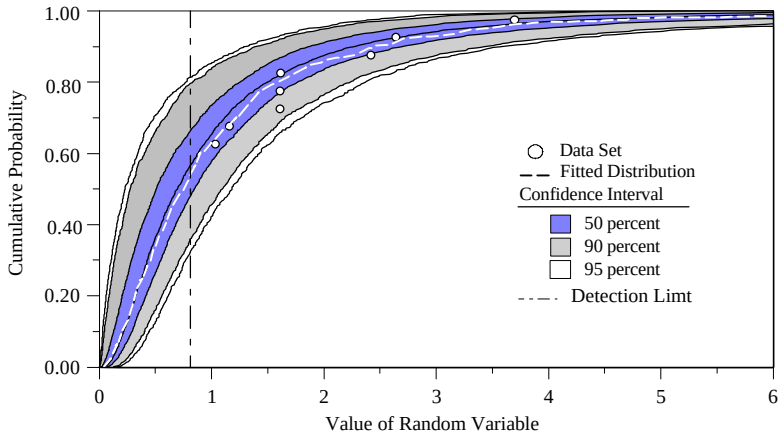
NC STATE UNIVERSITY

Results: Fitted Lognormal Distribution, 30% Censoring



NC STATE UNIVERSITY

Results: Fitted Lognormal Distribution, 60% Censoring



NC STATE UNIVERSITY

Results for Test Case 1: Gamma Distribution

Censoring percentage		0%	30%	60%
Number of non-detected data		0	6	12
Number of detected data		20	14	8
Total Data Points		20	20	20
Mean	Best estimate	1.07	1.06	1.10
	2.5^{th} percentile	0.64	0.65	0.61
	97.5^{th} percentile	1.74	1.69	1.71
	Width of 95% C.I.	1.10	1.04	1.10

NC STATE UNIVERSITY

Results for Test Case 1: Lognormal Distribution

Censoring percentage		0%	30%	60%
Number of non-detected data		0	6	12
Number of detected data		20	14	8
Total Data Points		20	20	20
Mean	Best estimate	1.01	1.00	0.97
	2.5^{th} percentile	0.65	0.63	0.51
	97.5^{th} percentile	1.51	1.53	1.52
	Width of 95% C.I.	0.86	0.90	1.01

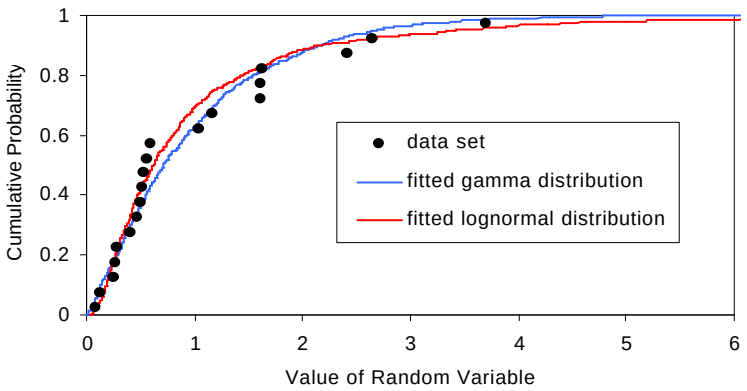
NC STATE UNIVERSITY

Results of Relative Uncertainty for Case 1

- For gamma distribution with 30% censoring, the relative uncertainty in mean is approximately –39% to +60%
- For lognormal distribution with 30% censoring, the relative uncertainty in mean is approximately –37% to +53%

NC STATE UNIVERSITY

Results: Comparison of Gamma and Lognormal, No Censoring for Case 1



Mean: data = 1.02, gamma = 1.07, lognormal = 1.01

NC STATE UNIVERSITY

Comparison of Mean Estimated by MLE and Conventional Approaches for Case 1

Censoring	Detected Points Only	Replacing Censored Data With			MLE	
		Zero	DL/2	DL	Gamma	Lognormal
0%	1.02				1.07	1.01
30%	1.35	0.95	1.01	1.08	1.06	1.00
60%	1.97	0.79	1.03	1.28	1.10	0.97

No estimate of uncertainty is available

NC STATE UNIVERSITY

Test Case 2

- 20 synthetic data points were generated from a gamma distribution (mean = 1 and standard deviation = 1)
- assign 0%, 30% and 60% censoring
- Gamma distribution were fit to the censored empirical bootstrap samples

NC STATE UNIVERSITY

Comparison of Mean Estimated by MLE and Conventional Approaches for Case 2

Censoring	Detected Points Only	Replacing Censored Data With			MLE
		Zero	DL/2	DL	
0%	1.05				1.04
30%	1.41	0.99	1.04	1.08	1.02
60%	2.04	0.81	1.11	1.23	1.00

No estimate of uncertainty is available

NC STATE UNIVERSITY

Uncertainty Results from MLE/Bootstrap Method for Test Case 2

Censoring Percentage		0%	30%	60%
Mean	Best estimate	1.04	1.02	1.00
	2.5^{th} percentile	0.62	0.60	0.51
	97.5^{th} percentile	1.56	1.55	1.58
	Width of 95% C.I.	0.94	0.95	1.06

- For 0% censoring, the relative uncertainty in mean is approximately -40% to +50%
- For 30% censoring, the relative uncertainty in mean is approximately -41% to +52%
- For 60% censoring, the relative uncertainty in mean is approximately -49% to +58%

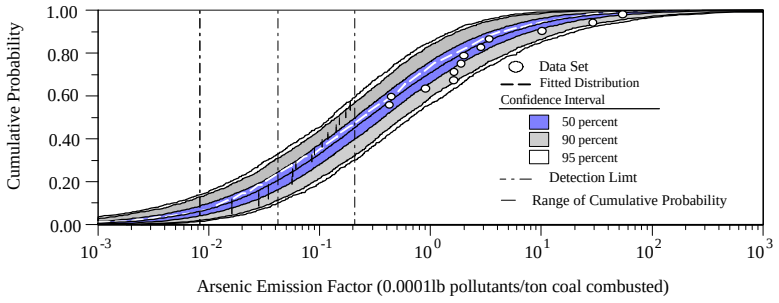
NC STATE UNIVERSITY

Example Case of Arsenic Emission Factor

- Case study: arsenic emission factor from coal combustion source
- 29 data points including 3 censored values
- Each censored data point has a different detection limit
- Some detected data values are less than some detection limits
- There is uncertainty regarding the empirical cumulative probability of such detected data values

NC STATE UNIVERSITY

Results of Example Case: Lognormal Distribution Fitted to Censored Data



- The 95 percent confidence interval for the mean is –91% to 264% of the mean value (8.2×10^{-4} lb arsenic / ton coal combusted)

NC STATE UNIVERSITY

Conclusions

- MLE is an asymptotically unbiased method for estimating the mean of censored data
- Successfully applied MLE to multiply censored data
- Successfully demonstrated quantification of uncertainty in the mean of censored data based upon bootstrap simulation
- Estimated variability and uncertainty in censored part of the distribution
- If mean is above the the detection limit(s), the uncertainty of the mean is not very sensitive to variation in the detection limit(s).

NC STATE UNIVERSITY

Acknowledgements

- The authors acknowledge the support of the Science to Achieve Results (STAR) grants program of the U.S. Environmental Protection Agency, which funded this work under Grant No. R826790. Although the research described in this presentation has been funded wholly or in part by the U.S. EPA, this presentation has not been subject to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred

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

NC STATE UNIVERSITY

Junyu (Allen) Zheng, Ph.D
H. Christopher Frey, Ph.D

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

NC STATE UNIVERSITY

Overview: Mixture Distribution

- Motivations and definition
- Methodologies for quantifying variability and uncertainty based upon mixture distributions
- Properties of quantification of variability and uncertainty with respect to variation in sample size, mixing weight and separation between components
- Example case study

NC STATE UNIVERSITY

Motivation: Mixture Distribution

- Single component distributions might not well describe the variation in a quantity
- Population distribution of a random variable is a mixture of distributions
- The use of single component distributions that are poor fits to data could potentially lead to bias in variability and uncertainty analysis.

NC STATE UNIVERSITY

Definition: Mixture Distributions

$$f(x) = w_1 f_1(x|\theta_1) + w_2 f_2(x|\theta_2) + \cdots + w_k f_k(x|\theta_k)$$

With $w_j > 0 \quad for \quad j = 1, ..., k$

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

Where

f(x) Probability density function for a mixture model
$f_j(x|\theta_j)$ Probability density function (PDF) for a component
w_J The mixing weight
θ_J Vector of parameters for a component

- Presently focus on two component Mixture Lognormal Distributions $f(x) = wf_1(x) + (1-w)f_2(x)$

NC STATE UNIVERSITY

Parameter Estimation: Mixture Distribution

- Maximum Likelihood Estimation (MLE)
  - MLE is widely used due to its relative efficiency and generality

$$L = \sum_{i=1}^{n} \ln[f(x_i|w,\mu,\sigma)] = \sum_{i=1}^{n} \ln\left\{\sum_{j=1}^{k} w_j f_j(x_i|\mu_j\sigma_j)\right\}$$

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
μ_j, σ_j: The parameters in the j^{th} component in a mixture distribution

NC STATE UNIVERSITY

Parameter Estimation: Mixture Distribution

- Procedures to find an approximate solution of the likelihood function in the MLE
  - EM algorithm
  - Newton-like Methods
  - Nonlinear optimization method
- Nonlinear optimization method was chosen
  - Straightforward
  - Does not require calculation of derivatives of

NC STATE UNIVERSITY

Quantification of Uncertainty and Variability: Mixture Distribution

- Specify mixture distribution: specifying w_i, θ_i
- Represent the mixture distribution as an empirical CDF
- Use empirical CDF as assumed population distribution for parametric simulation
- Bootstrap simulation
  - Method for generating bootstrap samples
  - Methods for forming confidence intervals

NC STATE UNIVERSITY

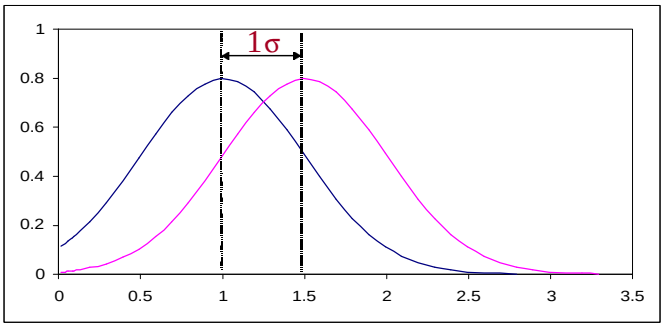
Quantification of Uncertainty and Variability: Mixture Distribution

- Method for generating bootstrap samples
  - Sampling algorithm based upon the empirical distribution was used
- Methods for forming confidence intervals
  - Percentile Method
    - » Easy to use
    - » Only first-order accurate
  - BC_a Method (Bias Correction and Acceleration)
    - » transformation respecting
    - » second-order accurate
    - » heavy computation load

NC STATE UNIVERSITY

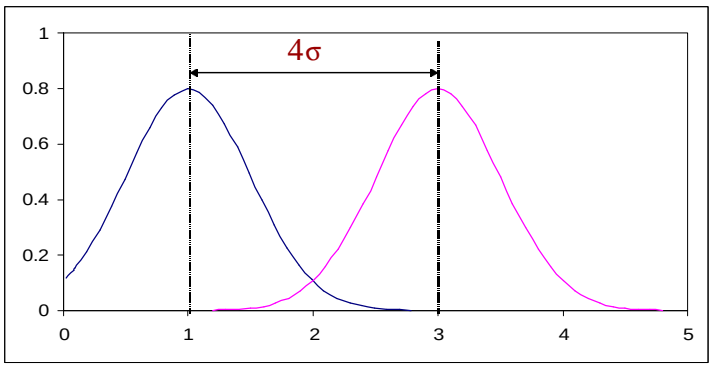
Properties: Study Design

- Sample size: 25, 50 and 100
- Mixing weight: 0.1, 0.3 and 0.5
- Separation between components: 1σ, 2σ, 4σ, 10σ



NC STATE UNIVERSITY

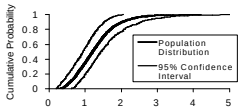
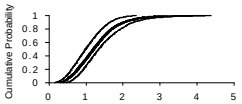
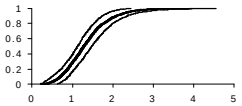
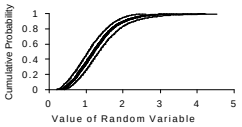
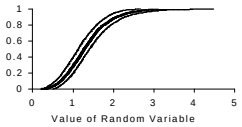
Study Design (Cont'd)



108 synthetic datasets which cover the variation in mixing weight, sample size and separation were

NC STATE UNIVERSITY

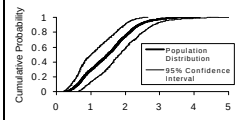
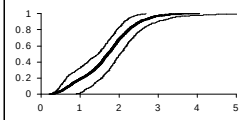
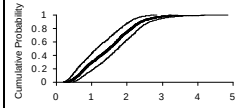
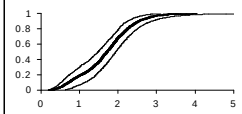
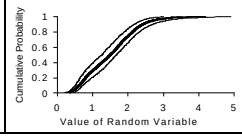
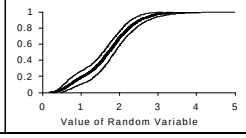
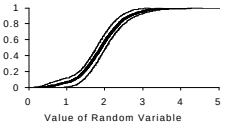
Separation=1 Standard Deviation

Sample Value	Weight=0.5	Weight=0.3	Weight=0.1
25		Failed	Failed
50			Failed
100			Failed Value of Random Variable

95 Percent Confidence Intervals of Cumulative Distributions of Two Component Lognormal Distributions Fitted to a Mixture Population Distributions (μ1=1.0, μ2=1.5, σ1=σ2=0.5) Based on Bootstrap Simulation (B=500)

NC STATE UNIVERSITY

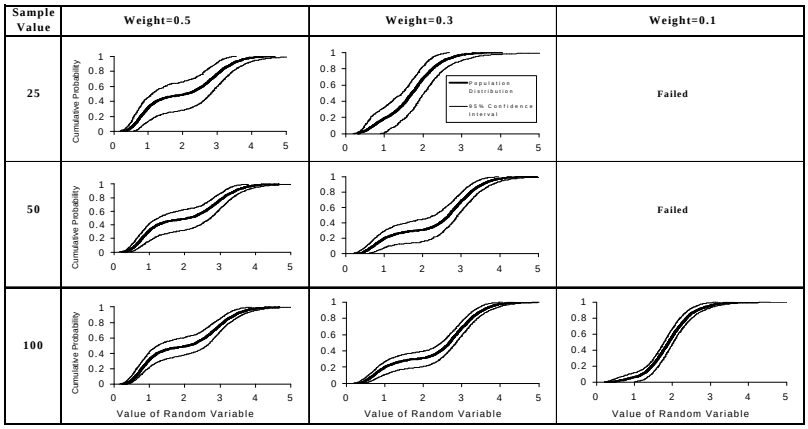
Separation=2 Standard Deviation

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

95 Percent Confidence Intervals of Cumulative Distributions of Two Component Lognormal Distributions Fitted to a Mixture Population Distributions (μ1=1.0, μ2=2.0, σ1=σ2=0.5) Based on Bootstrap Simulation (B=500)

NC STATE UNIVERSITY

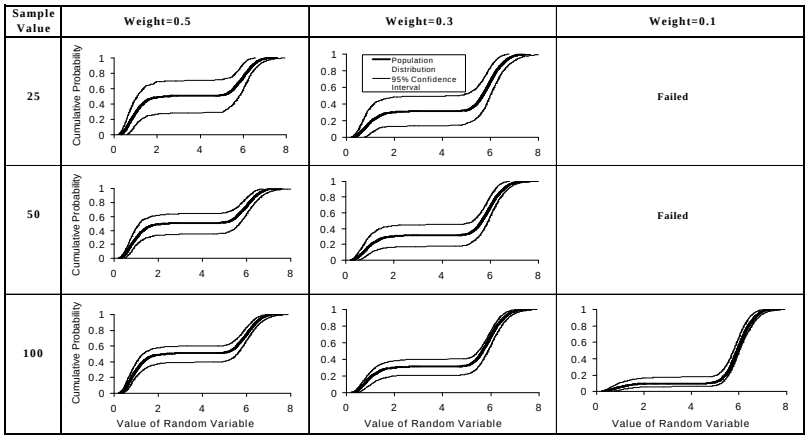
Separation=4 Standard Deviation



95 Percent Confidence Intervals of Cumulative Distributions of Two Component Lognormal Distributions Fitted to a Mixture Population Distributions (μ1=1.0, μ σ σ

NC STATE UNIVERSITY

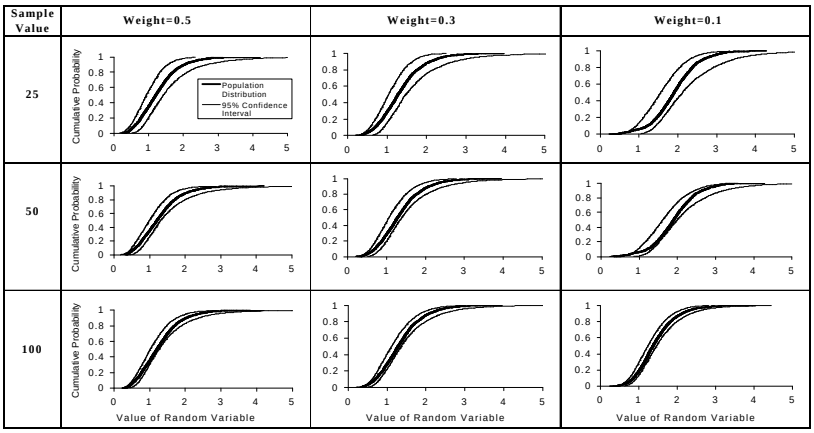
Separation=10 Standard Deviation



95 Percent Confidence Intervals of Cumulative Distributions of Two Component Lognormal Distributions Fitted to a Mixture Population Distributions (μ1=1.0, μ2=6.0, σ1=σ2=0.5) Based on Bootstrap Simulation (B=500)

NC STATE UNIVERSITY

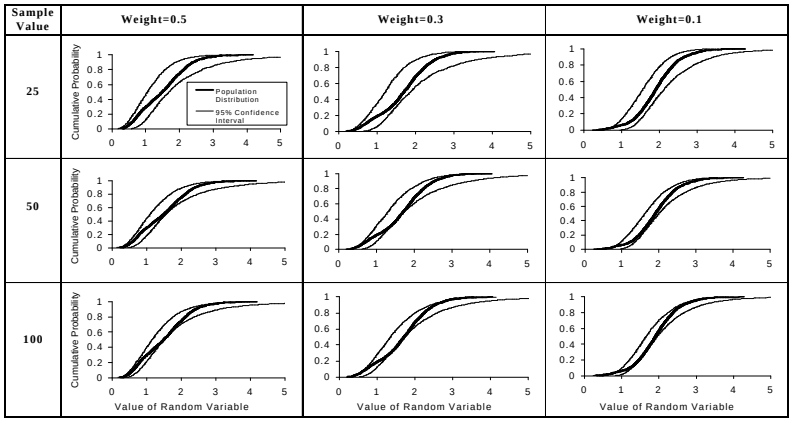
Use of Single Distribution: Separation=1σ



95 Percent Confidence Intervals of Cumulative Distributions of a Single Distribution Fitted to a Mixture Population Distributions (μ1=1.0, μ2=1.5, σ1=σ2=0.5) Based on Bootstrap Simulation (B=500)

NC STATE UNIVERSITY

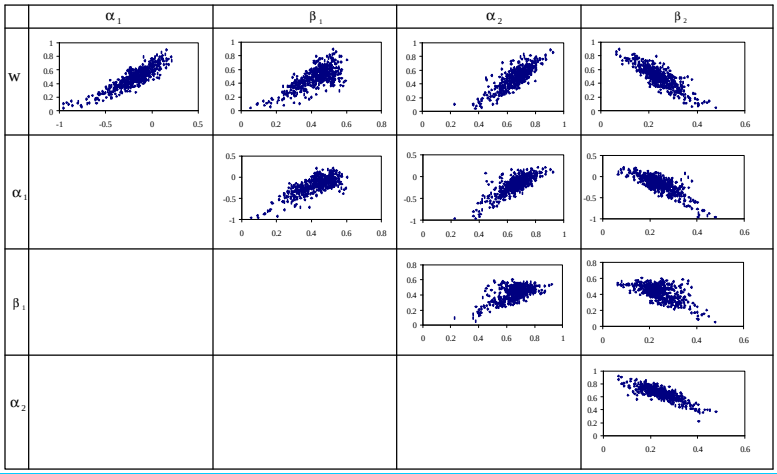
Use of Single Distribution: Separation=2σ



95 Percent Confidence Intervals of Cumulative Distributions of a Single Distribution Fitted to a Mixture Population Distributions (μ1=1.0, μ2=2.0, σ1=σ2=0.5) Based on Bootstrap Simulation (B=500)

NC STATE UNIVERSITY

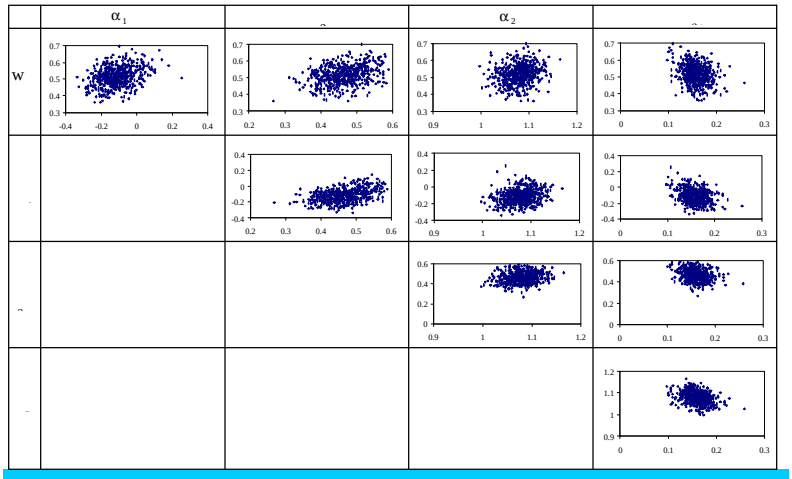
Parameter Dependency (Lig ht ly Separated)



Scatter Plots of Bootstrap Simulation Results (B=500) for Two Component Lognormal Distributions (μ1=1.0, μ2=2.0, σ1=σ2=0.5)

NC STATE UNIVERSITY

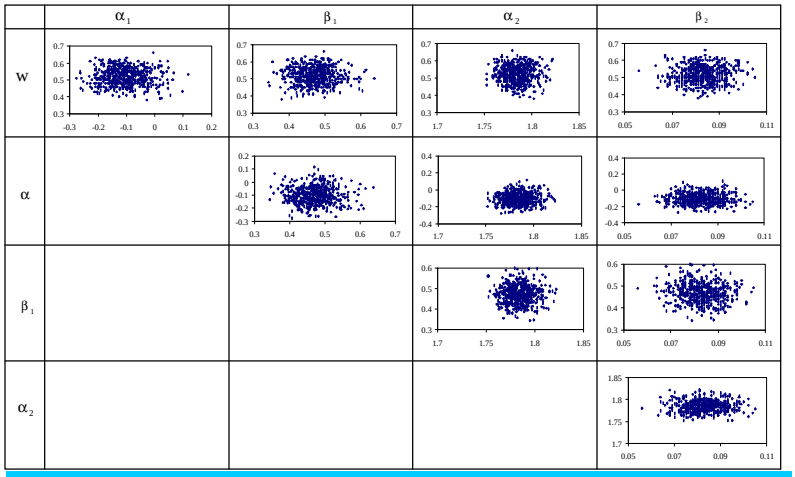
Parameter Dependency (Moderately Separated)



Scatter Plots of Bootstrap Simulation Results (B=500) for Two Component Lognormal Distributions (μ1=1.0, μ2=3.0, σ1=σ2=0.5)

NC STATE UNIVERSITY

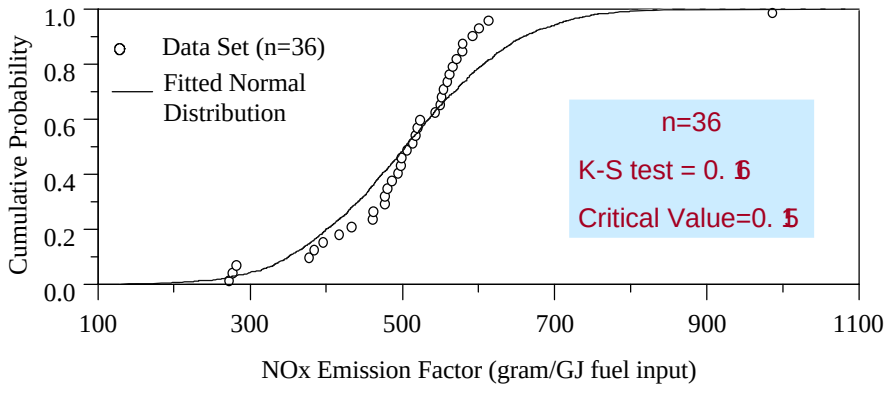
Parameter Dependency (Highly Separated)



Scatter Plots of Bootstrap Simulation Results (B=500) for Two Component Lognormal Distributions (μ1=1.0, μ2=3.0, σ1=σ2=0.5)

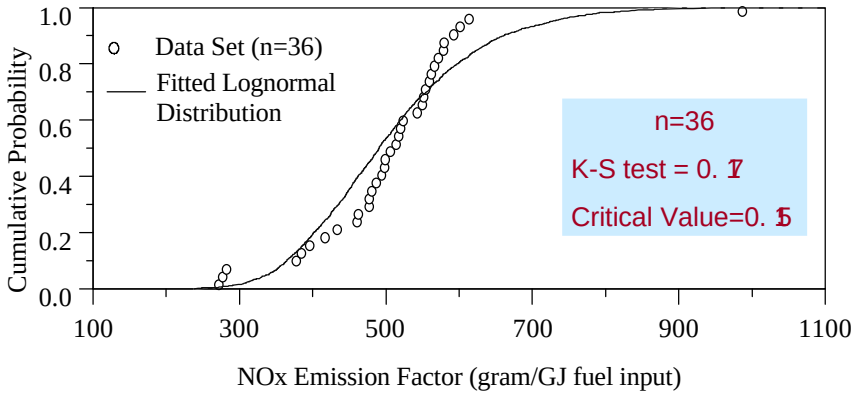
NC STATE UNIVERSITY

Example Case Study: Fitted Normal Distribution



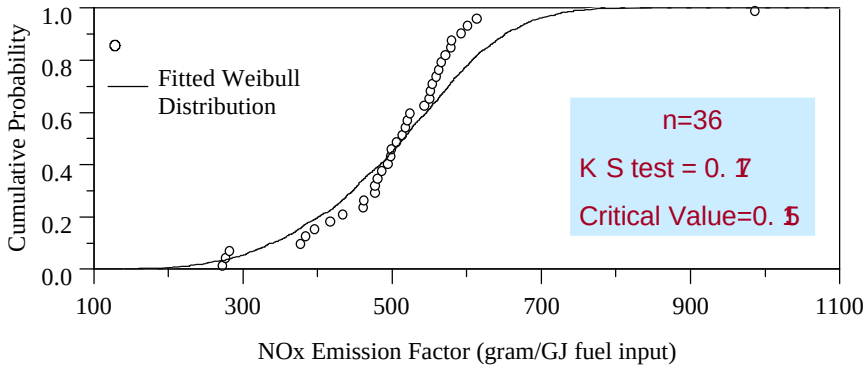
NC STATE UNIVERSITY

Example Case Study: Fitted Lognormal

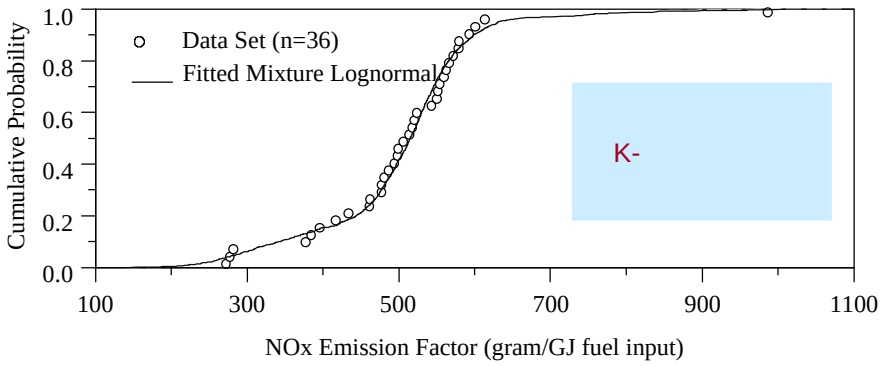


NC STATE UNIVERSITY

Example Case Study: Fitted Weibull Distribution



NC STATE UNIVERSITY



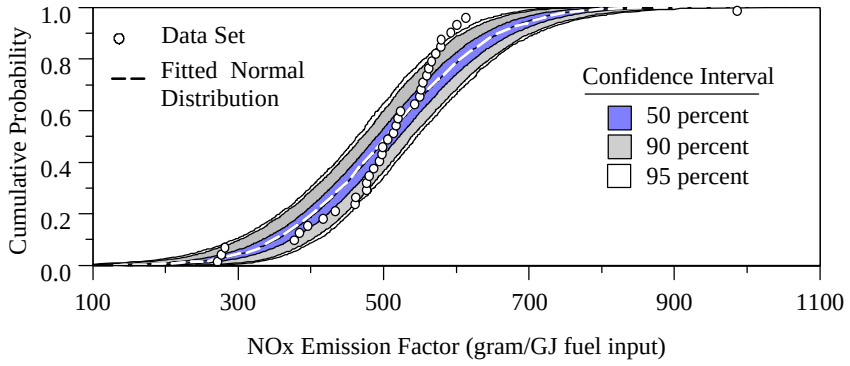
Weight = 0.352

1st Component: Mean of ln(x)=6.064, Std. Dev. of ln(x)=0.345

2nd Component: Mean of ln(x)=6.269, Std. Dev. of ln(x)=0.0847

NC STATE UNIVERSITY

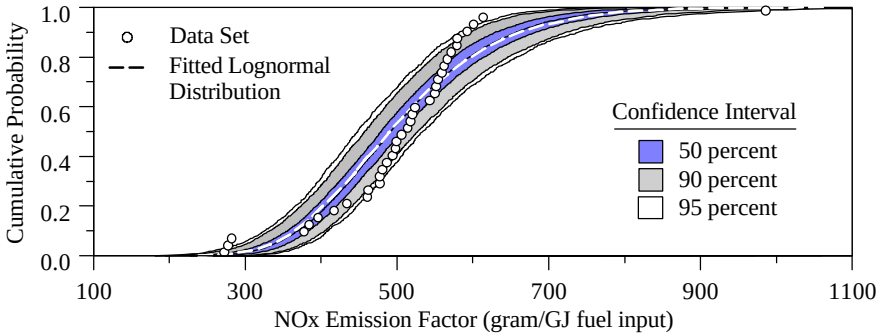
Probability Band (Normal Distribution)



Percent of data in 50% CI: 35%

NC STATE UNIVERSITY

Probability Band (Lognormal Distribution)

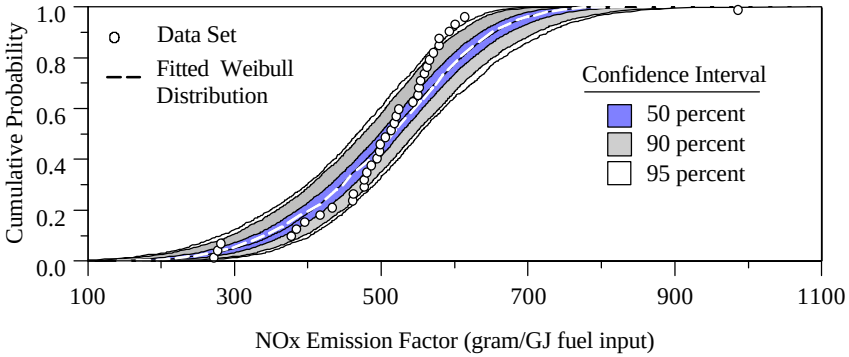


Percent of data in 50% CI 30%

Percent of data in 95% CI 83%

NC STATE UNIVERSITY

Probability Band (Weibull Distribution)

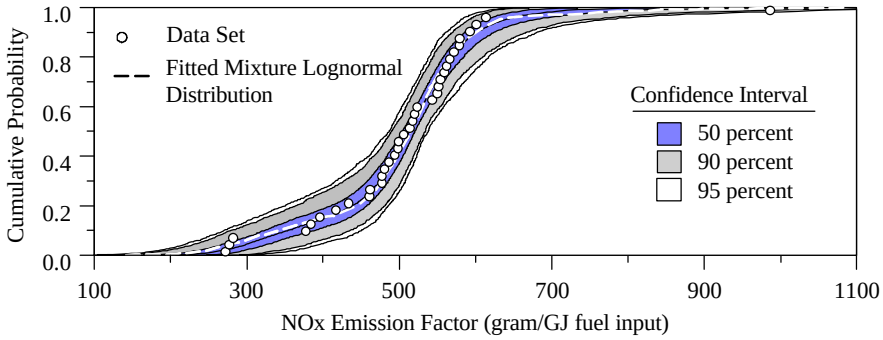


Percent of data in 50% CI 35%

Percent of data in 95% CI 86%

NC STATE UNIVERSITY

Probability Band (Mixture Lognormal Distribution)



Percent of data in 50% CI: 92%

Percent of data in 95% CI: 100%

NC STATE UNIVERSITY

Uncertainty in the Mean (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 10 simulations)

NC STATE UNIVERSITY

Uncertainty in the 95% Percentile of Variability (B=500)

Distribution Type	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]
Normal	635 [633,638]	701 [699, 702]	768 [765,772]
Lognormal	627 [623,633]	713 [711, 714]	813 [808,819]
Weibull	627 [624,631]	692 [691, 693]	778 [770,785]

[a]: Two component mixture lognormal distributions
[L, U][b]: Lower bound and upper bound (based upon 10 simulations)

The observed data at the 95% percentile of empirical CDF is 612.6

NC STATE UNIVERSITY

Summary and Conclusion

- A method was developed to quantify variability and uncertainty based upon mixture distribution
- Bootstrap simulation results tend to be more stable normally for larger sample size
- When two components are well separated, the stability and accuracy of quantification of variability and uncertainty are improved
- Typically, there is greater uncertainty regarding percentile of mixture distributions coinciding with the separated region
- When two components are not well separated, a single distribution may often be a better choice because it has fewer parameters and higher numerical stability

NC STATE UNIVERSITY

Summary and Conclusion (Cont'd)

- Dependencies may exist in sampling distributions of parameters of mixtures and are influenced by the amount of separation between the components
- The case study results indicate that a mixture lognormal distribution is a better fit to the selected case compared to single distributions
- The mixture distribution has potential to yield more efficient statistical estimates.
- 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

NC STATE UNIVERSITY

Quantification of Variability and Uncertainty with Known Measurement Error

NC STATE UNIVERSITY

Junyu (Allen) Zheng, Ph.D
H. Christopher Frey, Ph.D

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

NC STATE UNIVERSITY

Overview: Measurement Error

- Motivation
- Measurement error and uncertainty
- Classification of measurement error
- Measurement error models
- Error free data construction
- Quantification of variability and uncertainty with measurement error
- Properties of solutions for variability and uncertainty via a case study

NC STATE UNIVERSITY

Motivation

- Measurement errors affects all statistical analysis, both formal and informal because it causes the probability distribution that generates the observed data to deviate from that which generates unobservable, error free data (Chesher, 1991)
- Potentially brings bias into variability and uncertainty analysis

NC STATE UNIVERSITY

Measurement Error and Uncertainty

- Measurement Error
  - The deviation of the result of measurement from the true value of the measurable quantity (Dieck,1992)
- Uncertainty of Measurement
  - An interval within which a true value of a measurement lies within a given probability (Rabinovich, 1999)

NC STATE UNIVERSITY

Classification of Measurement Error

- Causes of Error (Rabinovich, 1999)
  - Methodological error
  - Instrument error
  - Personal error
- Properties of Error (Ellis, 1966; Barford, 1985)
  - Systematic error
  - Random error

NC STATE UNIVERSITY

Measurement Error Models

- Additive Model

$$Z_i = X_i + e_i$$

Where: Z_i = error contaminated data, observed data

X_i = error free data (true value)

e_i =Measurement error, often be assumed as a normal distribution with mean 0

- Multiplicative model

$$z_i = X_i e_i^*$$

- Multiplicative model can be log-transformed into the additive model

NC STATE UNIVERSITY

Quantification of Variability and Uncertainty: Error Free Data Construction

- Deconvolution Method
  - Assumption
    - » Known measurement error
    - » Additive measurement error models

$$_{z} = \int_{x} - \, dx$$

$f_e(e)$ = PDF for the measurement error, often assumed as a normal distribution
$f_z(z)$ = PDF for the observed data set
$f_x(x)$ = PDF for the error free data

  - Potential Problems
    - » Complicated mathematical inferences and computations
    - » Not a common probability distribution for the $f_x(x)$, potential difficulties in sampling algorithms

NC STATE UNIVERSITY

Quantification of Variability and Uncertainty: Error Free Data Construction

- Alternative Approach
  - Assumption
    - » Known measurement error and variance of the measurement error is less than variance of the observed dataset

$$\hat{X}_i = cz_i + (1-c)\bar{z}$$

Where:
$\hat{X}_i$ = Estimated error free data
z_i = Observed or error contaminated data
$\bar{z}$ = Sample mean of error contaminated or observed data
c = Constant; can be found by:

$$c = \sqrt{\frac{S_z^2 - \sigma_e^2}{S_z^2}} \qquad (S_z^2 > \sigma_e^2)$$

NC STATE UNIVERSITY

Quantification of Variability and Uncertainty wit h Known Measurement Error

- Bootstrap pair technique

$$\mathbf{Z_i}^{*}=(\mathbf{x_i}^{*}, \mathbf{e_i}^{*}) = x_{i,j} + e_{i,j} \qquad (i=1,2,..\ B;\ j=1,\ 2,\ldots n)$$

Where:

B = The number of bootstrap replication

n = The sample size of a dataset

$x_{i,j}$ = A random sample from a distribution describing error free data

$e_{i,j}$ = A random sample from a distribution describing measurement error

- Two-dimensional framework for characterizing uncertainty due to random sampling error
- Incorporation of the uncertainty from measurement error

NC STATE UNIVERSITY

Case Study

- Purpose
  - To demonstrate the use of the methods
  - To investigate the effect of the size of measurement errors on the variability and uncertainty estimates
- Study design

$$r = \frac{\sigma_e}{\sigma_{Total}}$$

Where:

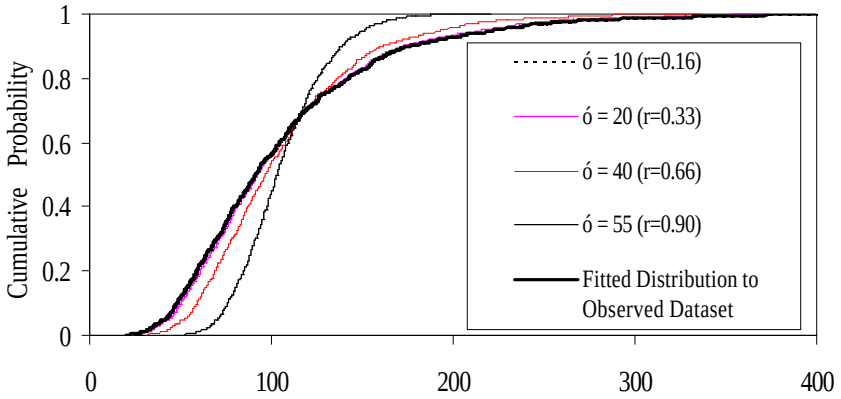
σ_e = Standard deviation of measurement error

σ_{Total} = Standard deviation of the observed error-contaminated data set

  - A synthetic dataset with mean of 107.3, standard deviation of 60.9 and sample size of 25
  - Measurement error models: N(0, 10)(r=0.16), N(20)(r=0.33), N(40)(r=0.66), N(55) (r=0.90)

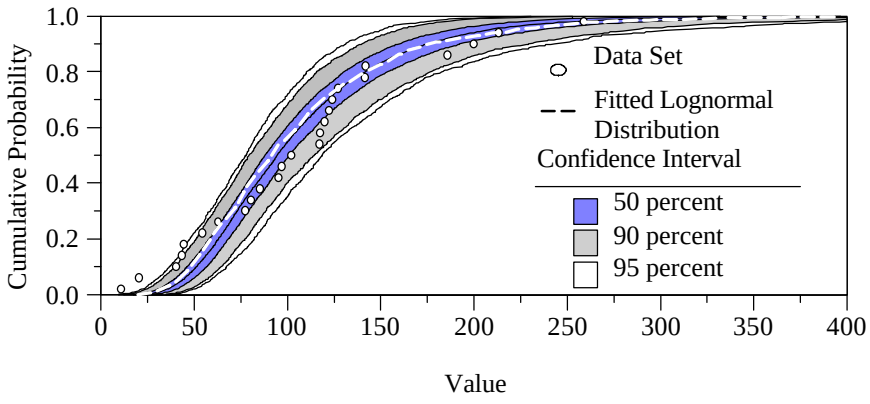
NC STATE UNIVERSITY

Fitted Distributions to the Error Free Datasets and the Observed Dataset



NC STATE UNIVERSITY

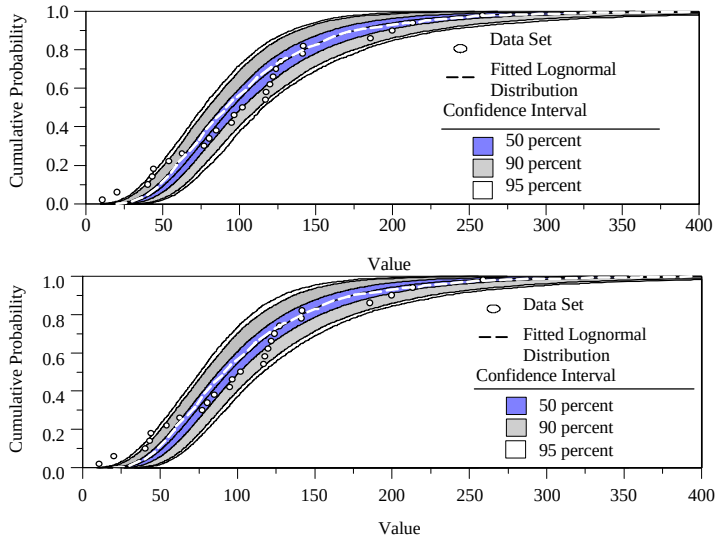
Probability Band: Observed Data Set



Random sample error and measurement error are not separated

NC STATE UNIVERSITY

Probability Band for Measurement Error Model: N(0.0, 10.0)

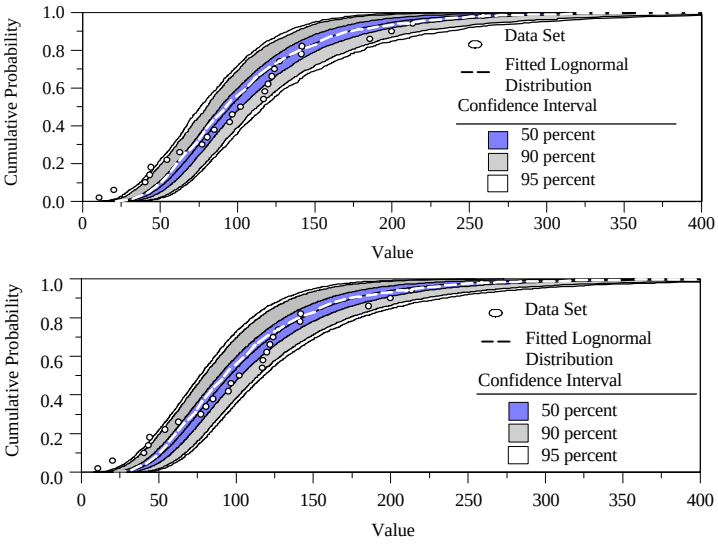


Uncertainty from measurement error is not included

Uncertainty from measurement error is included

NC STATE UNIVERSITY

Probability Band for Measurement Error Model: N(0.0, 20.0)

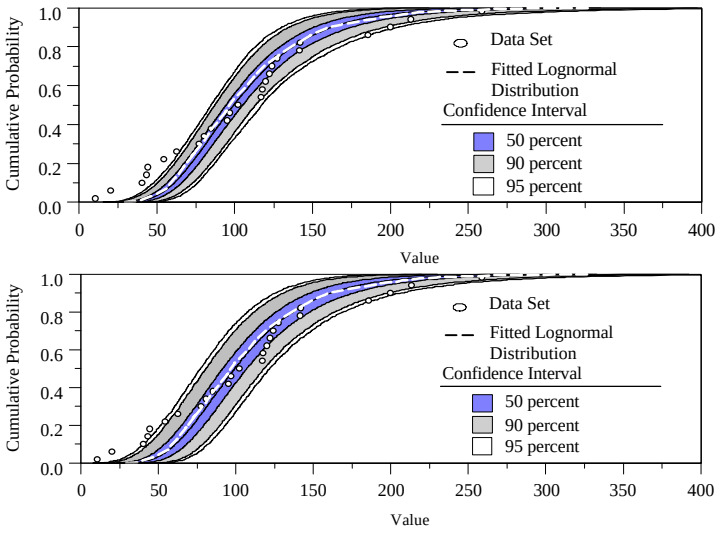


Uncertainty from measurement error is not included

Uncertainty from measurement error is included

NC STATE UNIVERSITY

Probability Band for Measurement Error Model: N(0.0, 40.0)

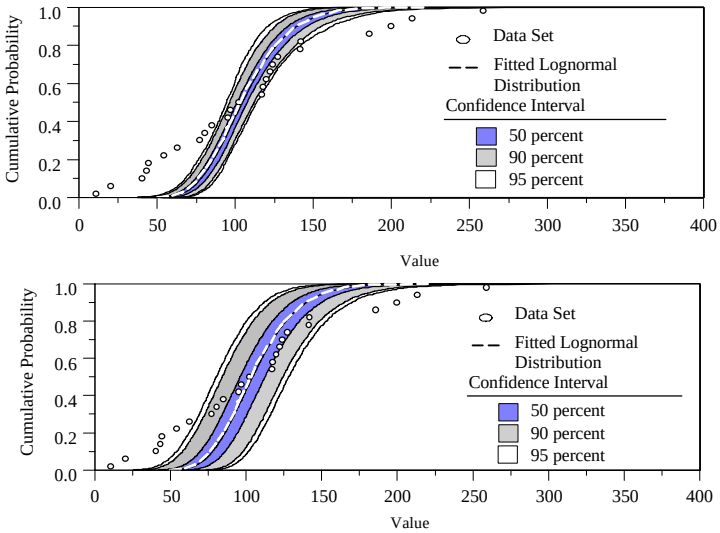


Uncertainty from measurement error is not included

Uncertainty from measurement error is included

NC STATE UNIVERSITY

Probability Band for Measurement Error Model: N(0.0, 55.0)

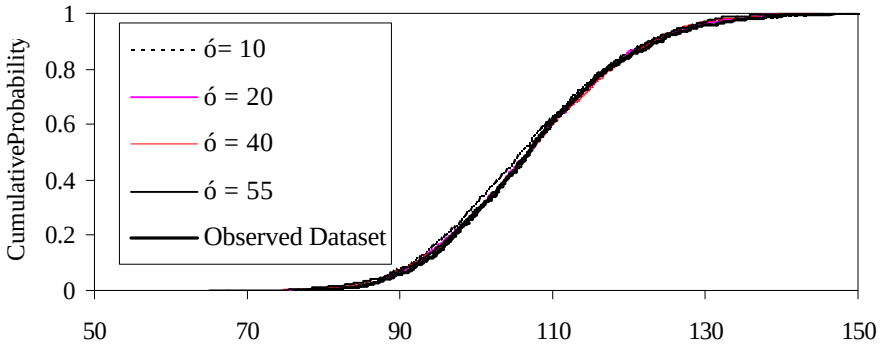


Uncertainty from measurement error is not included

Uncertainty from measurement error is included

NC STATE UNIVERSITY

Sampling Distributions for the Mean under Different Measurement Error Models.



NC STATE UNIVERSITY

Uncertainty in the Mean under Different Measurement Error Models.

Measurement Error model	Analysis Method [a]	Confidence Intervals for Mean (Random Sampling Error only)			Confidence Intervals for Mean (Both Random Sampling and Measurement Error)		
		2.5%	Mean	97.5%	2.5%	Mean	97.5%
N(0.0, 0.0)*	Analytic	83.4	107.3	132.4	83.4	107.3	132.4
	Numerical	85.8	107.3	132.0	85.8	107.3	132.0
N(0.0, 10.0)	Analytic	83.7	107.3	132.1	83.4	107.3	132.4
	Numerical	86.4	107.2	132.3	85.8	107.2	132.4
N(0.0, 20.0)	Analytic	84.7	107.3	131.1	83.4	107.3	132.4
	Numerical	87.4	107.5	131.4	84.9	107.2	132.3
N(0.0, 40.0)	Analytic	89.3	107.3	126.3	83.4	107.3	132.4
	Numerical	90.8	107.3	126.3	84.1	107.3	132.2
N(0.0, 55.0)	Analytic	97.0	107.3	118.1	83.4	107.3	132.4
	Numerical	97.5	107.3	117.7	84.2	107.4	130.7

Note: The results listed here are the average values of 10 different simulations for each case.
*****: Random sampling error and measurement error are not separated for this case.
[a]: Analytical solutions are based upon central limit theorem; numerical solutions are estimated from bootstrap simulation.

NC STATE UNIVERSITY

Summary and Conclusion

- A method is developed for constructing an error free data set based on the observed data set
- Demonstrates methods for improving the characterization of variability and uncertainty if there are known measurement errors in environmental data
- There exist substantial bias in the estimates of true variability if measurement error is substantial
- Uncertainty will be underestimated if uncertainty arising from measurement error is subtracted not characterized

NC STATE UNIVERSITY

Summary and Conclusion (Cont'd)

- No substantial difference among 95% confidence intervals and sampling distributions for the mean for the observed data set and the error free data sets if the contribution from measurement error to the total uncertainty is considered
- To get an unbiased estimate of true variability, it is necessary to separate measurement error from the observed variability.

NC STATE UNIVERSITY

Outline

- Introduction
- Quantification of variability
- Quantification of uncertainty in a single component distribution
- Introduction to AuvTool
- Quantification of variability and uncertainty in censored datasets
- Characterization of variability and uncertainty based upon mixture distributions
- Characterization of variability and uncertainty with known measurement error

NC STATE UNIVERSITY

Discussion and Questions