# Fitting Second-Order Parametric Distributions to Data
# Using Maximum Likelihood Estimation

David E. Burmaster
Alceon Corporation
PO Box 382669 Harvard Square Station
Cambridge, MA 02238-2669
tel: 617-864-4300; fax: 617-864-9954
deb@Alceon.com

Kimberly M. Thompson
Harvard University School of Public Health
718 Huntington Avenue
Boston, MA 02115
tel: 617-432-4345; fax: 617-432-0190
KimT@hsph.harvard.edu

Keywords

variability, uncertainty, second-order random variables, maximum likelihood estimation

## Abstract

Second-order random variables, i.e., parametric random variables with uncertain parameters, give risk assessors a way to distinguish and represent both the variability and the uncertainty in an exposure variable. In this manuscript, we explore ways to fit second-order random variables to data using maximum likelihood estimation (MLE).

## 1.0 Introduction

When estimating exposures to chemicals in the environment, risk assessors need a way to represent both the *variability* and the *uncertainty* in the exposure variables (see, e.g., NAS, 1994; NCRP, 1996; Burmaster & Wilson, 1996). In a probabilistic exposure assessment, a risk assessor may use a second-order probability distribution to represent the *variability* and the *uncertainty* in one or more of the exposure variables (Bogen, 1990; MacIntosh et al, 1994; McKone, 1994; Frey & Rhodes, 1996; Hattis & Barlow, 1996; Price et al, 1996). Although other authors have used either professional judgment (e.g, Hoffman & Hammonds, 1994; NCRP, 1996; Barry, 1996; Cohen et al, 1996) or the bootstrap method (e.g., Frey, 1996) to develop second-order random variables used in calculations and in "two-dimensional" Monte Carlo simulations, we show how to use the method of maximum likelihood estimation (MLE) to fit second-order distributions to data. In a "two-dimensional" Monte Carlo simulation, some or all of the input variables and all of the output variables are represented by second-order random variables (Burmaster & Wilson, 1996).

Sir Ronald A. Fisher developed the method of maximum likelihood estimation (MLE) as a powerful, general purpose method for fitting a parametric distribution to data. The general idea is to choose an estimator for the parameter(s) in a distribution as that function of the sample observations (i.e., data) which will, when substituted into the distribution, make the probability of the sample a maximum (paraphrased from Keeping, 1995). He later generalized the idea to develop joint confidence regions for the parameter(s), an idea that was further generalized to the profile likelihood method for marginal distributions for parameter(s).

In this manuscript, we apply some of the standard MLE techniques to an original and perturbed version of each of three data sets to show how to fit a second-order random variable to data. Before using the MLE methods, we first select an appropriate family of distributions through exploratory data analysis. In each of these six cases, MLE methods readily determine the joint and marginal distributions for the data sets in a way that can be easily used in "two-dimensional" Monte Carlo simulations.

## 2.0 Three Data Sets, Each with a Perturbation

In this manuscript, we consider three data sets, each in its original form and also in a perturbed form. The first two data sets are synthetic (and unitless), but the third data set comprises clinical measurements of body weights (kg) for a random sample of adult women between the ages of 18 and 40 years. For each of the three data sets, we investigate the original data and the perturbed data to see how the perturbation changes the second-order distributions fit by maximum likelihood estimation (MLE).

Data Set 1 Original (DS1o in Table 1), a synthetic data set, contains 19 positive values drawn randomly from a LogNormal distribution of the form exp[Normal($\mu$, $\sigma$)] with $\mu = 2$ and $\sigma = 1$ and then rounded to the nearest integer. The arithmetic mean of the parent distribution equals $\exp[\mu + 0.5\ \sigma^2] = 12.2$, approximately, and the arithmetic mean of this sample equals 14, exactly. When tested by the Wilk-Shapiro (W-S) test for Normality (Madansky, 1988), the natural logarithms of these 19 data points do not fail using standard criteria (p-value = 0.1492).

Data Set 1 Perturbed (DS1p in Table 1), also a synthetic data set, contains 19 positive values. The first 18 values are the same as those in DS1o, but we have perturbed the largest value to change it from 101 in DS1o to 301 in DS1p. The arithmetic mean of these 19 values equals 24.5, approximately. When tested by the Wilk-Shapiro (W-S) test for Normality, the natural logarithms of these 19 data points do fail using standard criteria (p-value = 0.0096).

Data Set 2 Original (DS2o in Table 2), a synthetic data set, contains 25 positive values randomly drawn from a Beta distribution of the form Beta[$\alpha$, $\beta$] with $\alpha = 2$ and $\beta = 6$ and then rounded to two decimal places. The arithmetic mean of the parent distribution equals $(\frac{\alpha}{\alpha+\beta}) = 0.25$, exactly, and the arithmetic mean of this sample equals approximately 0.28.

Data Set 2 Perturbed (DS2p in Table 2), also a synthetic data set, contains 25 positive values. The first 24 values are the same as those in DS2o, but we have perturbed the largest value to change it from 0.61 in DS2o to 0.75 in DS2p. The arithmetic mean of these 25 values equals approximately 0.28.

Data Set 3 Original (DS3o in Table 3), a real data set, contains values of body weight (kg) measured for a random sample of 1,958 adult women in the United States between the ages of 18 and 40 years (Stern, 1996). In this instance, the original researchers reported only key percentiles of the data, not all the individual values. Data Set 3 Perturbed (DS3p in Table 3) differs from the original measurements in the sense that we consider how our inferences would change if the same percentile statistics arose from measuring the body weights of 979 women (i.e., the same percentiles for body weight, but from half as many subjects). Since the values in DS3o come from empirical measurements for a random sample of adult women, we do not know *a priori* if they can be fit by a parametric distribution; however, we do know that one or two LogNormal distributions have been fit successfully to empirical measurements for the body weights of another (larger) random sample of adult women between the ages of 18 and 75 years (Brainard & Burmaster, 1992). Since we know selected percentiles for DS3o and do not know the individual measurements in the data set, we will use a more general form of MLE methods to fit a second-order distribution to these "binned" data.

## 3.0 Overview of the MLE Approach

As detailed in the next sections, we fit second-order random variables to the original and the perturbed versions of each of the three data sets in five steps. In the first step, we use graphical methods from exploratory data analysis to see if a parametric distribution may reasonably fit the data. In the second step, we fit a first-order random variable, i.e, an ordinary random variable represented by parametric distribution with fixed parameters, to the data. In the third step, we develop and explore the likelihood function (and the loglikelihood function) for the data (see, for example: Mood et al, 1974; Edwards, 1992; Keeping, 1995). In the final step, we differentiate the loglikelihood function to develop and fit second-order random variables to the data (Cox & Snell, 1989; Ross, 1990). Although the MLE method is quite general, it is important to check the intermediate and final results using graphs and plots.

In the next sections, we denote (i) real variables and real functions with plain letters, (ii) first-order random variables with a single underscore, (ii) second-order random variables with a double underscore, and (iv) vectors or matrices in bold.

## 4.0 The Method Applied to Data Set 1 -- Original and Perturbed

We know *a priori* that these synthetic data come from a LogNormal distribution. Combining the first and second steps, we use LogNormal probability plots (Burmaster & Hull, 1996; D'Agostino & Stephens, 1986) to fit this LogNormal distribution to the original and the perturbed versions of DS1 (Aitchison & Brown, 1957; Crow & Shimizu, 1988):

$$\ln[\, \underline{X} \,] \sim \text{Normal}[\, \mu, \sigma \,] \qquad \text{Eqn 1}$$

which is equivalent to:

$$\underline{X} \sim \exp[\, \text{Normal}[\, \mu, \sigma \,] \,] \qquad \text{Eqn 1'}$$

where ln[ • ] represents the Napierian (or natural) logarithm function, exp[ • ] represents the exponential function, and Normal[ $\mu$, $\sigma$ ] represents the Normal or Gaussian distribution with mean $\mu$ and standard deviation $\sigma$ (with $\sigma > 0$). From the probability plots shown in Figures 1.1-O and -P, we find these point values for $\hat{\mu}$ and $\hat{\sigma}$ from the intercept and the slope, respectively, of the straight line fit to the plot using ordinary least-squares regression:

| | Data Set 1 Original | Data Set 1 Perturbed |
|---|---|---|
| $\hat{\mu}$ | 2.014 | 2.072 |
| $\hat{\sigma}$ | 0.992 | 1.110 |
| adj$R^2$ | 0.922 | 0.845 |

Based on standard methods in logarithmic space, the *probability distribution* for drawing a single random value from the model in Eqn 1 is (Evans et al, 1993):

$$p[\ \ln[\ \underline{X}\ ]\ |\ \mu, \sigma\ ] \quad = \quad \frac{1}{\sigma\sqrt{2\,\pi}} \bullet \exp\left[\ -\frac{1}{2} \bullet \frac{(\ln(x) - \mu)^2}{\sigma^2}\ \right] \qquad \text{Eqn 2}$$

and the *likelihood function* for a single, randomly drawn sample, $x_i$, is:

$$p[\ \mu, \sigma\ |\ \ln(\ x_i\ )\ ] \quad = \quad \frac{1}{\sigma\sqrt{2\,\pi}} \bullet \exp\left[\ -\frac{1}{2} \bullet \frac{(\ln(x_i) - \mu)^2}{\sigma^2}\ \right] \qquad \text{Eqn 3}$$

In this framework, the *probability* of drawing N independent random samples is:

$$\text{Probability} \quad = \quad \Pi\ [\ p[\ \ln(x_i)\ |\ \mu, \sigma\ ]\ ] \qquad \text{Eqn 4}$$

and the *likelihood function* for the N independent random samples is:

$$\text{Likelihood} \quad = \quad \Pi\ [\ p[\ \ln(x_i)\ |\ \mu, \sigma\ ]\ ] \qquad \text{Eqn 5}$$

The *loglikelihood function* for the N independent random samples is a function of $\mu$ and $\sigma$:

$$\text{LogLikelihood} \quad = \quad \Sigma\ [\ \ln\ [\ p[\ \mu, \sigma\ |\ \ln(x_i)\ ]\ ]\ ]$$

$$J[\ \mu, \sigma\ ] \quad = \quad \Sigma\left[\ -\frac{1}{2}\ln[2\,\pi] - \ln[\sigma] - \frac{1}{2} \bullet \frac{(\ln(x_i) - \mu)^2}{\sigma^2}\ \right]$$

$$= \quad -\frac{N}{2}\ln[2\,\pi] - N\ln[\sigma] - \frac{1}{2} \bullet \Sigma\left[\ \frac{(\ln(x_i) - \mu)^2}{\sigma^2}\ \right] \qquad \text{Eqn 6}$$

The values of $\mu$ and $\sigma$ that maximize the loglikelihood function for the sample are called the MLE estimates $\hat{\mu}$ and $\hat{\sigma}$; each is a point value.

| | Data Set 1<br>Original | Data Set 1<br>Perturbed |
|---|---|---|
| $\hat{\mu}$ | 2.014 | 2.072 |

| | | |
|---|---|---|
| $\hat{\sigma}$ | 0.997 | 1.163 |
| max J | -26.9 | -29.8 |

In this example, the loglikelihood function has a single maximum at { $\hat{\mu}$, $\hat{\sigma}$ }. In Figures 1.2-O and -P, the dots near the center of the plot show the locations of { $\hat{\mu}$, $\hat{\sigma}$ }.

Again using standard methods (Mood et al, 1974; Cox & Snell, 1989; Edwards, 1992; Keeping, 1995), certain contours of the loglikelihood function define the joint confidence region for { $\mu$, $\sigma$ }. For example, the 95-percent joint confidence region is defined by this contour:

$$J[\mu, \sigma] \quad = \quad J[\hat{\mu}, \hat{\sigma}] - \frac{\chi^2_{0.05}}{2} \qquad \text{(with} \quad \text{df} = 2) \qquad \text{Eqn 7}$$

$$= \quad J[\hat{\mu}, \hat{\sigma}] - \frac{5.991}{2}$$

where $\chi^2_{0.05}$ refers to the ChiSquared distribution with two degrees of freedom (df = 2). Similarly, the 90-percent and 50-percent joint confidence regions follow similar contours with $\chi^2_{0.10}$ and $\chi^2_{0.50}$, respectively, substituted into Eqn 7 (each with df = 2). The solid lines Figures 1.2-O and -P show the 95-, 90- and 50-percent joint confidence regions as the largest, intermediate, smallest ovals, respectively (Wolfram, 1991; Wickham-Jones, 1994). Box & Tiao (1973, Chapter 2) present and discuss similar plots (and their corresponding marginal distributions) in an illuminating way.

Again using standard methods (Mood et al, 1974; Cox & Snell, 1989; Edwards, 1992; Keeping, 1995), the observed information matrix for the sample equals:

$$\mathbf{ObsInfo} \quad = \quad \begin{matrix} -\dfrac{\partial^2 J}{\partial \mu^2} & -\dfrac{\partial^2 J}{\partial \mu \partial \sigma} \\ -\dfrac{\partial^2 J}{\partial \sigma \partial \mu} & -\dfrac{\partial^2 J}{\partial \sigma^2} \end{matrix} \quad \hat{\mu}, \hat{\sigma} \qquad \text{Eqn 8}$$

and, under the standard Taylor series approximation and the standard regularity conditions (both met in this example), $\mu$ and $\sigma$ are distributed according to a MultiVariate Normal (MVN) distribution with this variance-covariance matrix: [EndNote 1].

$$\mathbf{\Sigma} \quad = \quad \text{Inverse}[\ \mathbf{ObsInfo}\ ] \qquad \text{Eqn 9}$$

$$= \begin{bmatrix} \mathrm{Var}[\,\mu\,] & \mathrm{Cov}[\,\mu\,\sigma\,] \\ \mathrm{Cov}[\,\sigma\,\mu\,] & \mathrm{Var}[\,\sigma\,] \end{bmatrix} \qquad \text{Eqn 10}$$

With the Taylor series approximation to the loglikelihood function, the approximations to the joint confidence regions are ellipses. For example, the ellipse that approximates the 95-percent joint confidence region for { μ, σ } is this contour of the MultiVariate Normal distribution (MVN) with df = 2:

$$\mathrm{MVN}[\,\underline{\mu}, \underline{\sigma}\,] \quad = \quad \frac{1}{2\,\pi\,\sqrt{\mathrm{Var}(\mu)\,\mathrm{Var}(\sigma)}\,\sqrt{1 - \dfrac{\mathrm{Cov}^2(\,\mu, \sigma\,)}{\mathrm{Var}(\,\mu\,)\,\mathrm{Var}(\,\sigma\,)}}} \bullet \exp\left[- \frac{\chi^2{}_{0.05}}{2}\right]$$

Eqn 10

In Eqn 10, $\chi^2{}_{0.05}$ refers to the ChiSquared distribution with two degrees of freedom (df = 2). Similarly, the 90-percent and 50-percent joint confidence regions follow similar ellipses with $\chi^2{}_{0.10}$ and $\chi^2{}_{0.50}$, respectively, substituted into Eqn 10 (each with df = 2).

Applying these methods to the original and the perturbed versions of DS1, we find that $\underline{\hat{\mu}}$ and $\underline{\hat{\sigma}}$ are each well approximated by Normal distributions with vanishing correlation.

| | Data Set 1 | Data Set 1 |
|---|---|---|
| | Original | Perturbed |
| $\underline{\hat{\mu}}$ | N(2.014, 0.229) | N(2.072, 0.267) |
| $\underline{\hat{\sigma}}$ | N(0.997, 0.162) | N(1.163, 0.189) |
| Corr[$\underline{\hat{\mu}}$, $\underline{\hat{\sigma}}$] | 0 | 0 |

and with the constraint $\underline{\hat{\sigma}} > 0$. Thus, we have now fit this second-order random variable to the data:

$$\ln[\,\underline{\underline{X}}\,] \sim \quad \mathrm{Normal}[\,\underline{\mu}, \underline{\sigma}\,] \qquad \text{Eqn 11}$$

which is equivalent to:

$$\underline{\underline{X}} \quad \sim \quad \exp[\,\mathrm{Normal}[\,\underline{\mu}, \underline{\sigma}\,]\,] \qquad \text{Eqn 11'}$$

The dashed lines Figures 1.2-O and -P show the 95-, 90- and 50-percent joint confidence regions as the largest, intermediate, smallest ellipses, respectively. In these figures, as expected, the joint confidence regions developed from the Taylor series approximation to the loglikelihood function (the ellipses shown with dashed lines) are similar to the joint confidence regions developed directly from the loglikelihood function (the ovals shown with solid lines). As the number of data points increases, the ellipses (dashed lines) and the ovals (solid lines) will converge.

Finally, in Figures 1.3-O and -P, the lines show the 95-percent confidence bands on the probability plot using the isopleths developed in Burmaster & Wilson (1996). [EndNote 2].

Discussion: As expected, the 19 data points in DS1o pass the Wilk-Shapiro test for Normality, are well fit on the probability plot in Figure 1.1-O, have small joint confidence regions (ellipses and ovals) in Figure 1.2-O, and all fall within the 95-percent confidence band in Figure 1.3-O. As estimated by different methods, the estimates $\{ \hat{\mu}, \hat{\sigma} \}$ are close to $\{ \mu = 2, \sigma = 1 \}$, the values used to synthesize the 19 original data points. In contrast, the 19 data points in DS1p do not pass the Wilk-Shapiro test for Normality, have a good-sized outlier on the probability plot in Figure 1.1-P, have larger joint confidence regions (ellipses and ovals) in Figure 1.2-P, and do not all fall within the 95-percent confidence band in Figure 1.3-P. As estimated by different methods, the estimates $\{ \hat{\mu}, \hat{\sigma} \}$ differ from $\{ \mu = 2, \sigma = 1 \}$ by about 4 percent and 20 percent, respectively. Overall, the perturbation applied to the largest datum propagates into larger relative changes in $\hat{\sigma}$ and $\underline{\hat{\sigma}}$ and into smaller relative changes in $\hat{\mu}$ and $\underline{\hat{\mu}}$.

## 5.0 The Method Applied to Data Set 2 - Original and Perturbed

We know *a priori* that these synthetic data come from a Beta distribution. Combining the first and second steps, we use empirical cumulative distributions (CDFs) to fit this Beta distribution to the original and the perturbed versions of DS2:

$$\underline{X} \quad \sim \quad \text{Beta}[\,\alpha, \beta\,] \qquad \text{Eqn 12}$$

On the empirical CDF plots shown in Figures 2.1-O and -P, we could have estimated $\hat{\alpha}$ and $\hat{\beta}$ by nonlinear regression.

Working in linear space, the *probability distribution* for drawing a single random value from the model in Eqn 12 is (Evans et al, 1993):

$$p[\, \underline{X} \,|\, \alpha, \beta \,] \quad = \quad \frac{x^{\alpha-1}\,(1 - x)^{\beta-1}}{\mathrm{BetaFn}(\, \alpha, \beta \,)} \qquad \text{Eqn 13}$$

where $\mathrm{BetaFn}(\, \alpha, \beta \,) = \int_0^1 u^{\,\alpha-1}\,(1\text{-}u)^{\,\beta-1}\,du$. *The likelihood function* for a single, randomly drawn sample, $x_i$, is:

$$p[\, \alpha, \beta \,|\, x_i \,] \quad = \quad \frac{x_i^{\alpha-1}\,(1 - x_i)^{\beta-1}}{\mathrm{BetaFn}(\, \alpha, \beta \,)} \qquad \text{Eqn 14}$$

In this framework, the *loglikelihood function* for N independent random samples is:

$$\mathrm{LogLikelihood} \quad = \quad \Sigma\, [\, \ln [\, p[\, \alpha, \beta \,|\, x_i \,] \,] \,]$$

$$J[\, \alpha, \beta \,]$$

$$= \quad \Sigma \Big[\, (\alpha\text{-}1) \ln[\, x_i \,] + (\beta\text{-}1) \ln[1 - x_i] - \ln[\mathrm{BetaFn}(\, \alpha, \beta \,)\,] \Big] \qquad \text{Eqn 15}$$

The values of α and β that maximize the loglikelihood function for the sample are called the MLE estimates $\hat{\alpha}$ and $\hat{\beta}$; each is a point value.

| | Data Set 2 | Data Set 2 |
|---|---|---|
| | Original | Perturbed |
| $\hat{\alpha}$ | 2.099 | 1.874 |
| $\hat{\beta}$ | 5.479 | 4.686 |
| max J | 13.2 | 11.6 |

In this example, the loglikelihood function has a single maximum at { $\hat{\alpha}$, $\hat{\beta}$ }. In Figures 2.2-O and -P, the dots near the center of the plot show the locations of { $\hat{\alpha}$, $\hat{\beta}$ }.

Using the same standard methods, certain contours of the loglikelihood function define the joint confidence region for { α, β }. For example, the 95-percent joint confidence region is defined by this contour:

$$J[\,\alpha, \beta\,] = J[\,\hat{\alpha}, \hat{\beta}\,] - \frac{\chi^{2}\ 0.05}{2} \qquad \text{(with df = 2)}$$

$$= J[\,\hat{\alpha}, \hat{\beta}\,] - \frac{5.991}{2} \qquad \text{Eqn 16}$$

and the 90- and 50-percent joint confidence regions follow similar contours with appropriate values of the ChiSquared distribution substituted in Eqn 16.

Under the same assumptions as the first example, α and β are distributed according to a MultiVariate Normal distribution (MVN) with the variance-covariance matrix equal to the inverse of the observed information matrix for the sample:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Var}[\,\alpha\,] & \mathrm{Cov}[\,\alpha\ \beta\,] \\ \mathrm{Cov}[\,\alpha\ \beta\,] & \mathrm{Var}[\,\beta\,] \end{bmatrix} \qquad \text{Eqn 17}$$

$$= \text{Inverse} \quad \begin{matrix} -\dfrac{\partial^2 J}{\partial \alpha^2} & -\dfrac{\partial^2 J}{\partial \alpha \partial \beta} \\ -\dfrac{\partial^2 J}{\partial \beta \partial \alpha} & -\dfrac{\partial^2 J}{\partial \beta^2} \end{matrix} \quad \hat{\alpha}, \hat{\beta} \qquad \text{Eqn 18}$$

With the Taylor series approximation to the loglikelihood function, the approximations to the joint confidence regions are ellipses. For example, the ellipse that approximates the 95-percent joint confidence region for { α, β } is this contour of the MultiVariate Normal distribution (MVN) with df = 2:

$$\mathrm{MVN}[\,\underline{\alpha}, \underline{\beta}\,] = \frac{1}{2\,\pi\,\sqrt{\mathrm{Var}(\alpha)\ \mathrm{Var}(\beta)}\ \sqrt{1 - \dfrac{\mathrm{Cov}^2(\,\alpha, \beta\,)}{\mathrm{Var}(\alpha)\ \mathrm{Var}(\beta)}}} \bullet \exp\left[- \frac{\chi^{2}\ 0.05}{2}\right] \qquad \text{Eqn 19}$$

Applying these methods to the original and the perturbed versions of DS2, we find that $\underline{\hat{\alpha}}$ and $\underline{\hat{\beta}}$ are each approximated by Normal distributions with large, positive correlation.

| | Data Set 2 | Data Set 2 |
|---|---|---|
| | Original | Perturbed |
| $\underline{\hat{\alpha}}$ | N(2.099, 0.555) | N(1.874, 0.493) |
| $\underline{\hat{\beta}}$ | N(5.479, 1.559) | N(4.686, 1.333) |

| | | |
|---|---|---|
| Corr($\underline{\hat{\alpha}}$, $\underline{\hat{\beta}}$) | 0.850 | 0.832 |

and with the two constraints $\underline{\hat{\alpha}}$ >1 and $\underline{\hat{\beta}}$ > 1. Note that Corr($\underline{\hat{\alpha}}$, $\underline{\hat{\beta}}$) is the correlation between $\underline{\hat{\alpha}}$ and $\underline{\hat{\beta}}$ . The results here correspond to this second-order random variable:

$$\underline{\underline{X}} \quad \sim \quad \text{Beta}[\,\underline{\alpha}, \underline{\beta}\,] \qquad \text{Eqn 20}$$

The dashed lines Figures 2.2-O and -P show the 95-, 90- and 50-percent joint confidence regions as the largest, intermediate, smallest ellipses, respectively. In these figures, as expected, the joint confidence regions developed from the Taylor series approximation to the loglikelihood function (the ellipses shown with dashed lines) are similar to the joint confidence regions developed directly from the loglikelihood function (the ovals shown in solid lines). As the number of data points increases, the ellipses and the ovals will again converge.

Finally, in Figures 2.3-O and -P, we approximate the 95-percent confidence bands on the empirical CDFs by plotting four Beta distributions in each with parameters chosen from the ends of the major and minor axes of the corresponding 95-percent ellipses.

Discussion: As expected, the 19 data points in DS2o are reasonably fit on the empirical CDF plot in Figure 2.1-O, but they have relatively large joint confidence regions (ellipses and ovals) in Figure 2.2-O. As estimated by different methods, the estimates { $\underline{\hat{\alpha}}$ , $\underline{\hat{\beta}}$ } are reasonably close to { $\alpha = 2$, $\beta = 6$ }, the values used to synthesize the 19 original data points. In contrast, the 19 data points in DS1p have a better fit on the empirical CDF in Figure 2.1-P, have smaller joint confidence regions (ellipses and ovals) in Figure 2.2-P, and also fall within the 95-percent confidence band in Figure 2.3-P. In this example, the perturbation applied to the largest datum propagates into a smaller joint confidence region for $\underline{\hat{\alpha}}$ and $\underline{\hat{\beta}}$ .

## 6.0 The Method Applied to Data Set 3 - Original and Perturbed

We do not know *a priori* that these measured data come from a parametric distribution. Letting $\underline{BW}$ denote the distribution of body weight, we follow the lead of Brainard and Burmaster (1992) by using LogNormal probability plots (Burmaster & Hull, 1996; D'Agostino & Stephens, 1986) to see if a LogNormal distribution can model the original and the perturbed versions of DS3:

$$\ln[\ \underline{BW}\ ] \quad \sim \quad \text{Normal}[\ \mu, \sigma\ ] \qquad \text{Eqn 21}$$

Since the points plotted in Figures 3.1-O and -P fall in a pattern with positive curvature, we immediately see that we may need to consider a more complex model, perhaps a mixture model (Brainard & Burmaster, 1992), for these data. However, since the curvature in each plot is small, we proceed to fit a single LogNormal distribution to the data. As shown in Figures 3.1-O and -P, we estimate $\hat{\mu}$ and $\hat{\sigma}$ in Eqn 21 by fitting a straight line to the plot using ordinary least-squares regression:

| | Data Set 3<br>Original | Data Set 3<br>Perturbed |
|---|---|---|
| $\hat{\mu}$ | 4.171 | same |
| $\hat{\sigma}$ | 0.222 | same |
| adjR$^2$ | 0.967 | same |

Since we do have the individual measured data, we must use a cumulative distribution function (CDF) to develop the loglikelihood function for these "binned" data (e.g., Tanner, 1996, p. 15). The CDF for the univariate Normal distribution is:

$$\text{CDF(Normal}(x \mid \mu, \sigma)) \quad = \quad \frac{1}{2}\left[\ 1 + \text{Erf}[\ \frac{x - \mu}{\sigma\sqrt{2}}\ ]\ \right] \qquad \text{Eqn 22}$$

where Erf[ • ] denotes the Error Function (Abramowitz & Stegun, 1964).

Given that the *probability* for <u>one</u> of the measurements for body weight (BW in kg) <u>in the bin</u> between $\{BW_{lo}, p_{lo}\}$ and $\{BW_{hi}, p_{hi}\}$ is:

$$\left[\ \text{CDF}[\ \ln[\ BW_{hi}] \mid \mu, \sigma] - \text{CDF}[\ \ln[\ BW_{lo}] \mid \mu, \sigma\ ]\ \right]$$

the *probability* for <u>all</u> of the measurements <u>in the bin</u> between $\{BW_{lo}, p_{lo}\}$ and $\{BW_{hi}, p_{hi}\}$ is:

$$\left[\ \text{CDF}[\ \ln[\ BW_{hi}] \mid \mu, \sigma] - \text{CDF}[\ \ln[\ BW_{lo}] \mid \mu, \sigma\ ]\ \right] \hat{}\ (\ N\ (p_{hi} - p_{lo})\ )$$

and the *loglikelihood function* for all of the measurements is:

$$J[\,\mu, \sigma\,]$$

$$= \quad N \bullet \sum (\, p_{hi} - p_{lo}\,) \bullet \ln\Big[\, \mathrm{CDF}[\,\mu, \sigma \,|\, \ln[BW_{hi}]\,] - \mathrm{CDF}[\,\mu, \sigma \,|\, \ln[\,BW_{lo}]\,]\,\Big] \qquad \text{Eqn 23}$$

The MLE estimates $\hat{\mu}$ and $\hat{\sigma}$ are:

| | Data Set 3<br>Original | Data Set 3<br>Perturbed |
|---|---|---|
| $\underline{\hat{\mu}}$ | 4.171 | same |
| $\underline{\hat{\sigma}}$ | 0.226 | same |
| max J | -7853. | -3926. |

and, by taking the inverse of the observed information matrix, the parameters for the fitted second-order distributions are:

| | Data Set 3<br>Original | Data Set 3<br>Perturbed |
|---|---|---|
| $\underline{\hat{\mu}}$ | N(4.171, 0.0051) | N(4.171, 0.0072) |
| $\underline{\hat{\sigma}}$ | N(0.226, 0.0037) | N(0.226, 0.0052) |
| Corr[$\underline{\hat{\mu}}$, $\underline{\hat{\sigma}}$] | 0 | 0 |

with the constraint $\underline{\hat{\sigma}} > 0$. As expected, the marginal distributions are uncorrelated.

Discussion: Even though Figures 3.1-O and -P suggest the need for a mixture model (see Figure 4 and related text in Brainard & Burmaster, 1992), we use MLE to fit second-order random variables to the original and perturbed data. As N decreases from 1,958 to 979, the joint confidence regions (in Figures 3.2-O and -P) grow in size but the 95-percent confidence bands (in Figures 3.3-O and -P) widen but do not enclose all the points in either plot (in Figures 3.3-O and -P).

At this point, we have two options to proceed. As a first option, we could acknowledge that the second-order LogNormal distributions just fit to the data do not capture the variability and the uncertainty in the data. We could then consider other parametric distributions or mixture models (Brainard & Burmaster, 1992). As a second option, we could use professional judgment to increase the amount of uncertainty in the second-order distribution. For example, when we increase the uncertainty in $\hat{\sigma}$ to Normal( 0.226, 0.0037 ), the 95-percent confidence bands in Figures 3.3-O and -P spread to enclose all the data points in each figure.

## 7.0 Discussion and Conclusions

The MLE method has many strengths: First, it works with every type of parametric distribution, including mixtures of parametric distributions. Second, it works with censored and/or binned data, e.g., measurements reported as "nondetect" with a stated detection limit. Third, it works with truncated distributions. Fourth, it produces joint confidence regions with the proper correlations among the parameters being estimated. Fifth, as the number of data points grows large, it converges asymptotically to Normal theory and produces joint confidence regions as ellipses. Sixth, with one, two, or three fitted parameters, it produces results that are easily visualized and used in "two-dimensional" Monte Carlo simulations.

The MLE method also has some limitations: First, when using the MLE method, the analyst must choose a family of (parametric) probability distributions. If the analyst chooses an inappropriate family of distributions, the MLE method will lead to suboptimal or erroneous results. If several families of distributions appear to fit the data reasonably well, the analyst must discriminate carefully among them. Thus, the analyst must use many lines of reasoning -- from computer visualizations to professional judgment -- when selecting the family of distributions. Second, the fitting procedures often require more power than is typically built into commercial spreadsheet programs. Third, it can be difficult to visualize the results if a distribution (or mixture of distributions) has more than three parameters.

Finally, we recognize the need to explore other approaches to fitting second-order random variables to data as well, especially: (i) methods based on Kolmogorov-Smirnov techniques (Bickel & Doksum; 1977); (ii) methods based on the bootstrap technique (Efron & Tibshirani, 1991 and 1993; Frey & Burmaster, 1997); and (iii) methods based on Bayesian techniques (Box & Tiao, 1973; Gelman et al, 1995; O'Ruanaidh & Fitzgerald, 1996; Sivia, 1996).

## Acknowledgments

We thank Ronald J. Bosch, Louis A. Cox, Jr., H. Christopher Frey, and Murray D. Smith for helpful suggestions during this research, and we thank Ronald J. Bosch for doing the Wilk-Shapiro tests (SAS, 1990). We also thank two anonymous reviewers for excellent suggestions for improving the manuscript.

Alceon Corporation funded this research.

## Dedication

We dedicate this manuscript to Frederick Mosteller.

## Trademarks

Mathematica ® is a registered trademark of Wolfram Research, Inc: http://www.wri.com
Alceon ® is a registered trademark of Alceon Corporation: http://www.alceon.com

## EndNotes

1. With $\boldsymbol{\xi}$ as a $k \times 1$ vector, the MultiVariate Normal distribution has this probability density function (Anderson, 1958; Rose & Smith, 1996):

$$\boldsymbol{\xi} \quad \sim \quad \mathrm{MVN}[\, \boldsymbol{\mu}, \boldsymbol{\Sigma} \,]$$

$$p[\, \boldsymbol{\xi} \,] \quad = \quad \frac{1}{|\boldsymbol{\Sigma}|^{1/2}\, (2\,\pi)^{k/2}} \bullet \exp\left[ -\frac{1}{2} \bullet (\boldsymbol{\xi} - \boldsymbol{\mu})^{T}\, \boldsymbol{\Sigma}^{-1}\, (\boldsymbol{\xi} - \boldsymbol{\mu}) \right]$$

2. The isopleths for the confidence bands are (Burmaster & Wilson, 1996):

$$\ln[\underline{\underline{X}}]\,[\, z_U \mid z_V \,] \quad \underset{=}{\bullet} \quad \mu_\mu + z_V \bullet \mu_\sigma + z_U \bullet \sqrt{\,(\sigma_\mu)^2 + (z_V \bullet \sigma_\sigma)^2}$$

where $\ln[\underline{\underline{X}}]\,[\, z_U \mid z_V \,]$ denotes the point value at $z_U$ conditional on $z_V$. In this equation, the symbol "$\underset{=}{\bullet}$" denotes "is approximately equal to." Thus, to approximate the point value for the 67th percentile of uncertainty on the 95th percentile of variability of the LogNormal distribution, first evaluate this equation with $z_V = 1.645$ and $z_U = 0.440$ and then exponentiate the result.

## References

Abramowitz & Stegun, 1964

Abramowitz, M. and I.A. Stegun, Eds, 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series Number 55, Issued June 1964, Tenth Printing with corrections in December 1972, US Government Printing Office, Washington, DC

Aitchison & Brown, 1957

Aitchison, J. and J.A.C. Brown, 1957, The Lognormal Distribution, Cambridge University Press, Cambridge, UK

Anderson, 1958

Anderson, T.W., 1958, An Introduction to Multivariate Statistical Analysis, Wiley, New York, NY

Barry, 1996

Barry, T.M., 1996, Distributions on a Budget, presented at US EPA's Workshop on Monte Carlo Analysis, 14 May 1996, New York, NY

Bickel & Doksum, 1977

Bickel, P.J. and K.A. Doksum, 1977, Mathematical Statistics, Basic Ideas and Selected Topics, Holden-Day, San Francisco, CA

Bogen, 1990

Bogen, K.T., 1990, Uncertainty in Environmental Risk Assessment, Garland Publishing, New York, NY

Box & Tiao, 1973

Box, G.E.P. and G.C. Tiao, 1973, Bayesian Inference in Statistical Inference, Wiley & Sons, New York, NY

Brainard & Burmaster, 1992

Brainard, J. and D.E. Burmaster, 1992, Bivariate Distributions for Height and Weight of Men and Women in the United States, Risk Analysis, 1992, Volume 12, Number 2, pp 267 - 275

Burmaster & Wilson, 1996

Burmaster, D.E. and A.M. Wilson, 1996, An Introduction to Second-Order Random Variables in Human Health Risk Assessment, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 892 - 919

Burmaster & Hull, 1996

Burmaster, D.E. and D.A. Hull, 1996, Using LogNormal Distributions and LogNormal Probability Plots in Probabilistic Risk Assessment, Human and Ecological Risk Assessment, in press

Cohen et al, 1996

Cohen, J.T., M.A. Lampson, and T.S. Bowers, 1996, The Use of Two-Stage Monte Carlo Simulation Techniques to Characterize Uncertainty and Variability in Risk Analysis, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 939 - 971

Cox & Snell, 1989

Cox, D.R. and E.J. Snell, 1989, Analysis of Binary Data, Second Edition, Chapman & Hall, London, UK

Crow & Shimizu, 1988

Crow, E.L. and K. Shimizu, Eds., 1988, Lognormal Distributions, Theory and Applications, Marcel Dekker, New York, NY

D'Agostino & Stephens, 1986

D'Agostino, R.B. and M.A. Stephens, 1986, Goodness-of-Fit Techniques, Marcel Dekker, New York, NY

Edwards, 1992

Edwards, A.W.F., 1992, Likelihood, John Hopkins University Press, Baltimore, MD

Efron & Tibshirani, 1993

Efron, B. and R.J. Tibshirani, 1993, An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability 57, Chapman & Hall, New York, NY

Efron & Tibshirani, 1991

Efron, B. and R.J. Tibshirani, 1991, Statistical Data Analysis in the Computer Age, Science, Volume 253, pp 390 - 395, 26 July 1991

Evans et al, 1993

Evans, M., N. Hastings, and B. Peacock, 1993, Statistical Distributions, Second Edition, John Wiley & Sons, New York, NY

Frey, 1996

Frey, H.C., 1996, Quantitative Techniques for Analysis of Variability and Uncertainty in Exposure Assessment, presented at US EPA's Workshop on Monte Carlo Analysis, 14 May 1996, New York, NY

Frey & Rhodes, 1996

Frey, H.C. and D.S. Rhodes, 1996, Characterizing, Simulating, and Analyzing Variability and Uncertainty: An Illustration of Methods Using an Air Toxics Emissions Example, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 762 - 797

Frey & Burmaster, 1997

Frey, H.C. and D.E. Burmaster, 1997, Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches, Risk Analysis, in preparation

Gelman et al, 1995

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin, 1995, Bayesian Data Analysis, Chapman & Hall, London, UK

Hattis & Barlow, 1996

Hattis, D. and K. Barlow, 1996, .Human Interindividual Variability in Cancer Risks -- Technical and Management Challenges, Human and Ecological Risk Assessment, Volume 2, Number 1, pp 194 - 220

Hoffman & Hammonds, 1994

Hoffman, F.O. and J.S. Hammonds, 1994, Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty due to Lack of Knowledge and Uncertainty due to Variability, Risk Analysis, Volume 7, Number 4, pp 707 - 712

Keeping, 1995

Keeping, E.S., 1995, Introduction to Statistical Inference, Dover, New York, NY

MacIntosh et al, 1994

MacIntosh, D.M., G.W. Suter, and F.O. Hoffman, 1994, Uses of Probabilistic Exposure Models in Ecological Risk Assessments of Contaminated Sites, Risk Analysis, Volume 14, Number 4, pp 405 - 420

Madansky, 1988

Madansky, A., 1988, Prescriptions for Working Statisticians, Springer-Verlag, New York, NY

McKone, 1994
McKone, T.E., 1994, Uncertainty and Variability in Human Exposures to Soil Contaminants through Home-Grown Food: A Monte Carlo assessment Risk Analysis, Volume 14, Number 4, pp 449 - 463

Mood et al, 1974
Mood, A.M., F.A. Graybill, and D.C. Boes, 1974, Introduction to the Theory of Statistics, Third Edition, McGraw Hill, New York, NY

NAS, 1994
National Academy of Sciences, 1994, Science and Judgment in Risk Assessment, National Academy Press, Washington, DC

NCRP, 1996
National Council on Radiation Protection and Measurement, 1996, A Guide for Uncertainty Analysis in Dose and Risk Assessments Related to Environmental Contamination, NCRP Commentary, Number 14, Bethesda, MD, 10 May 1996

O'Ruanaidh & Fitzgerald, 1996
O'Ruanaidh, J.K. and W.J. Fitzgerald, 1996, Numerical Bayesian Methods Applied to Signal Processing, Springer Verlag, New York, NY

Price et al, 1996
Price, P.S., S.H. Su, J.R. Harrington, and R.E. Keenan, 1996, Uncertainty and Variability in Indirect Exposure Assessments: An Analysis of Exposure to Tetrochlorodibenzo-p-dioxin from a Beef Consumption Pathway, Risk Analysis, Volume 16, Number 2, pp 263 - 277

Rose & Smith, 1996
Rose, C. and M.D. Smith, 1996, The Multivariate Normal Distribution, Mathematica Journal, Volume 6, pp 32 - 37, Winter 1996

Ross, 1990
Ross, G.J.S., 1990, Nonlinear Estimation, Springer-Verlag, New York, NY

SAS, 1990
SAS Institute, 1990, SAS/STAT User's Guide, Volume 2, Version 6, 4th Edition, Cary, NC

Sivia, 1996
Sivia, D.S., 1996, Data Analysis, A Bayesian Tutorial, Clarendon Press, Oxford, UK

Stern, 1996
Stern, A.H., 1996, NJ Department of Environmental Protection, personal communication with D.E. Burmaster

Tanner, 1996
Tanner, M.A., 1996, Tools for Statistical Inference, Springer-Verlag, New York, NY

Wickham-Jones, 1994
Wickham-Jones, T., 1994, Mathematica Graphics, Techniques & Applications, Springer-Verlag, Telos, Santa Clara, CA

Wolfram, 1991
Wolfram, S., 1991, Mathematica®, A System for Doing Mathematics by Computer, Second Edition, Addison- Wesley, Redwood City, CA

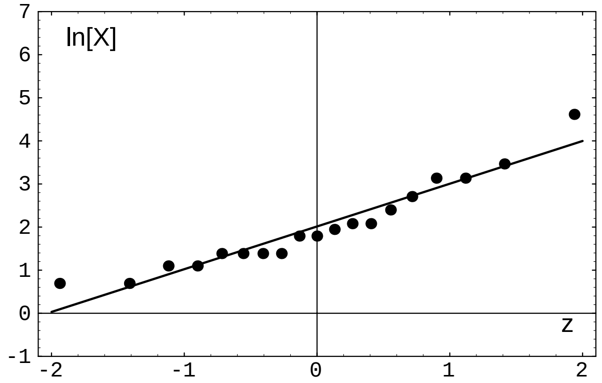

Figure 1.1-O: LogNormal Probability Plot, Original

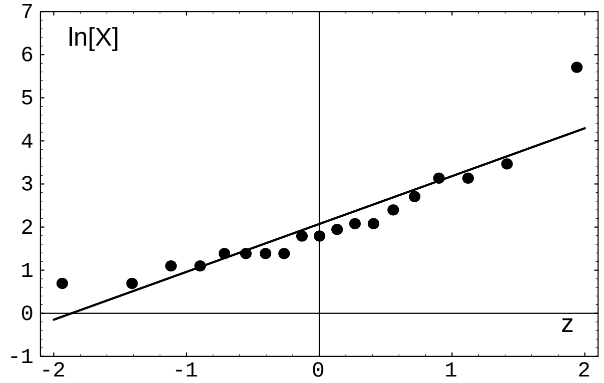

Figure 1.1-P: LogNormal Probability Plot, Perturbed

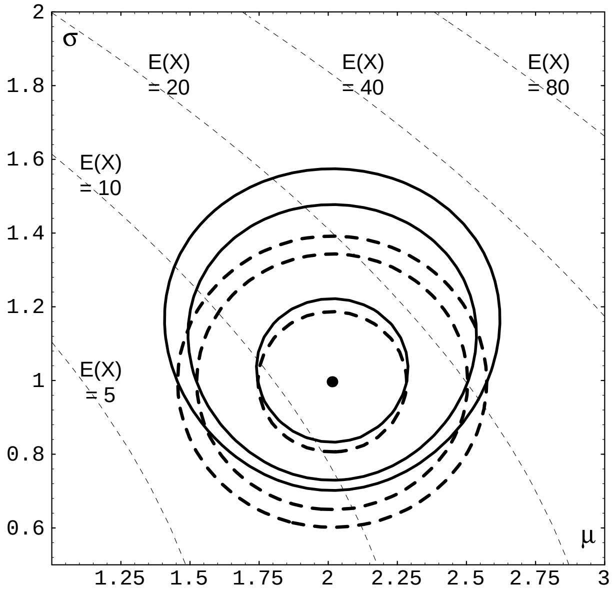

Figure 1.2-O: Joint Confidence Regions, Original

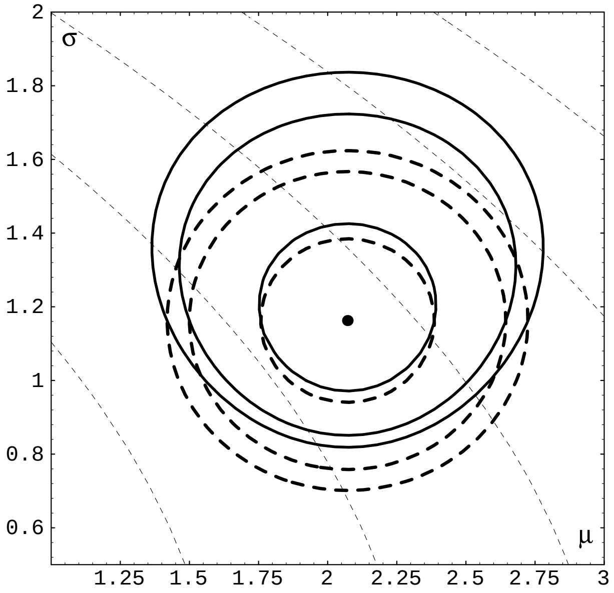

Figure 1.2-P: Joint Confidence Regions, Perturbed

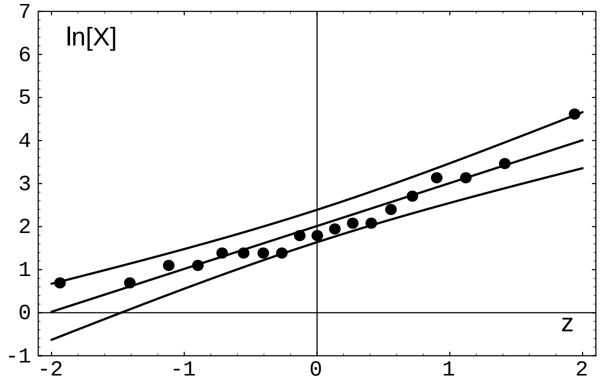

Figure 1.3-O: Confidence Bands, Original

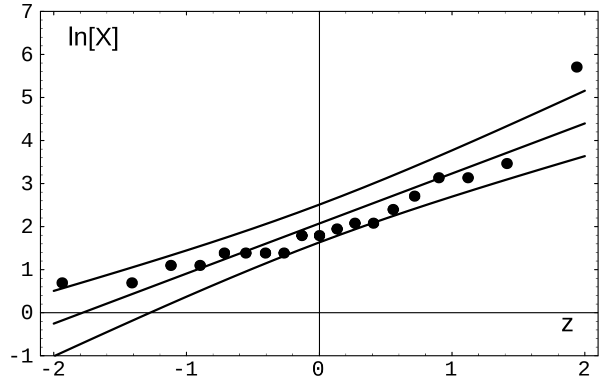

Figure 1.3-P: Confidence Bands, Perturbed

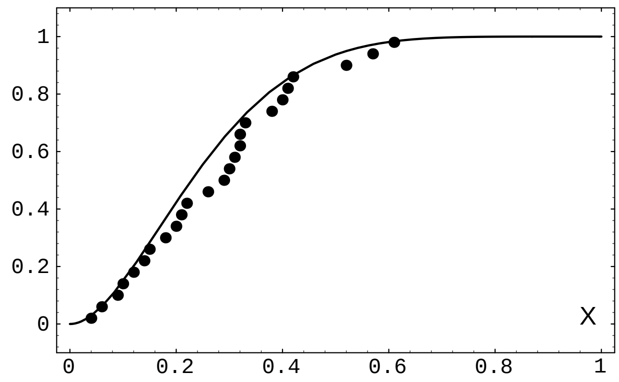

Figure 2.1-O: Cumulative Distribution, Original

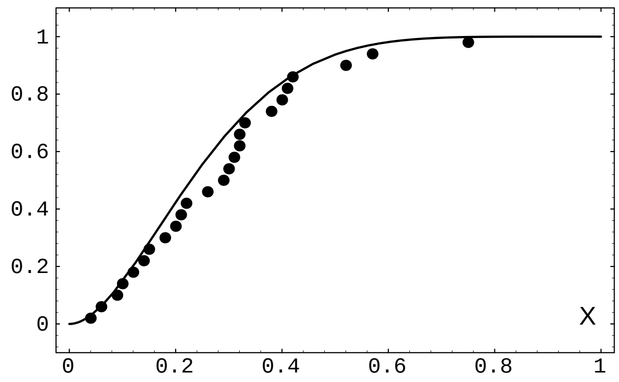

Figure 2.1-P: Cumulative Distribution, Perturbed

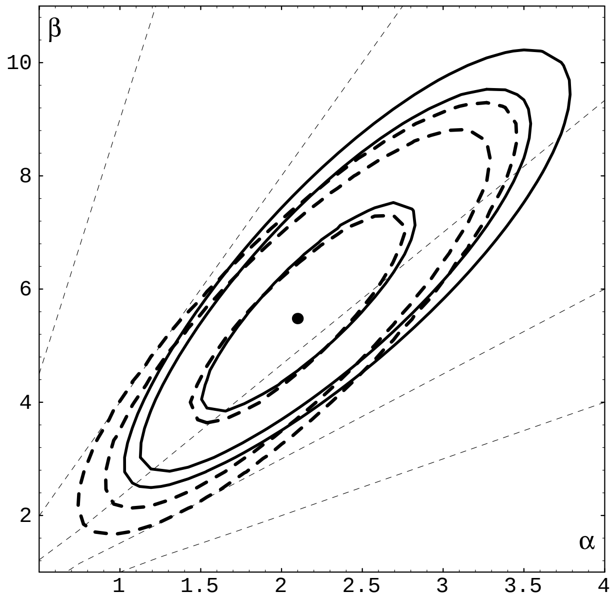

Figure 2.2-O: Joint Confidence Regions, Original

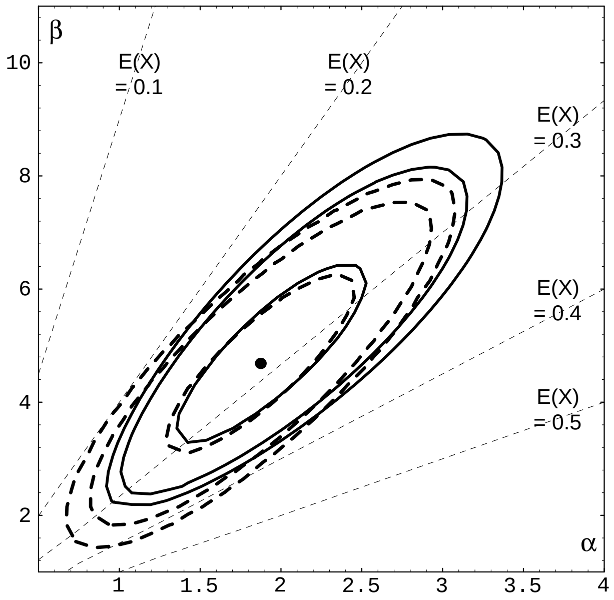

Figure 2.2-P: Joint Confidence Regions, Perturbed

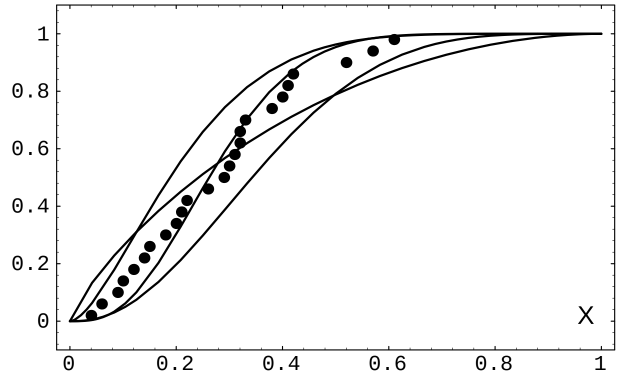

Figure 2.3-O: Confidence Bands, Original

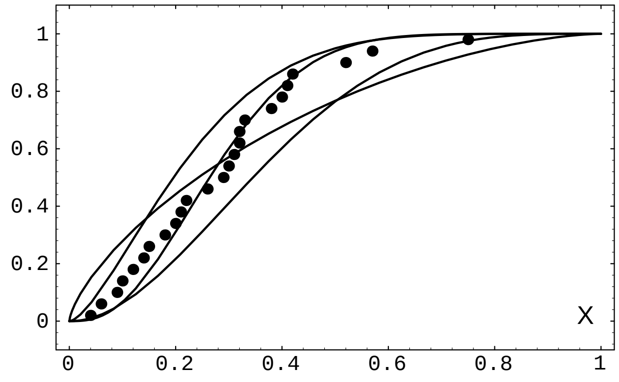

Figure 2.3-P: Confidence Bands, Perturbed

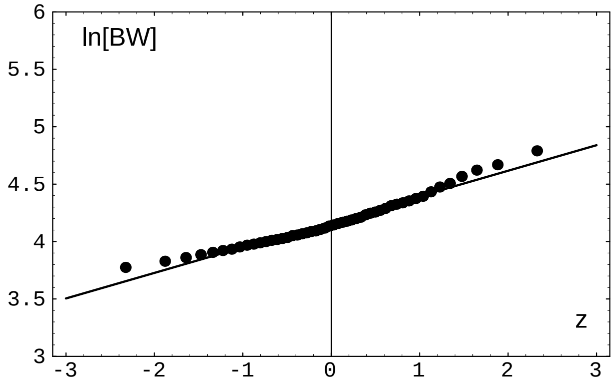

Figure 3.1-O: LogNormal Probabiliity Plot, n = 1,958

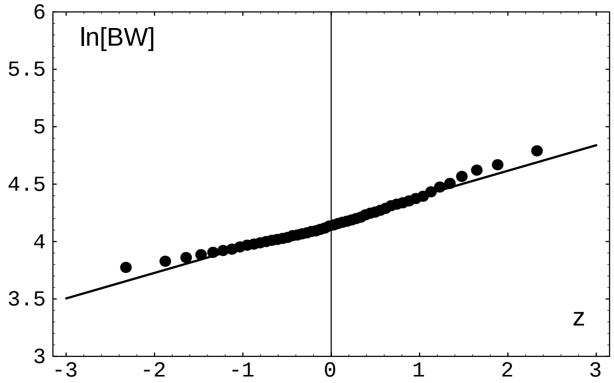

Figure 3.1-P: LogNormal Probability Plot, n = 979

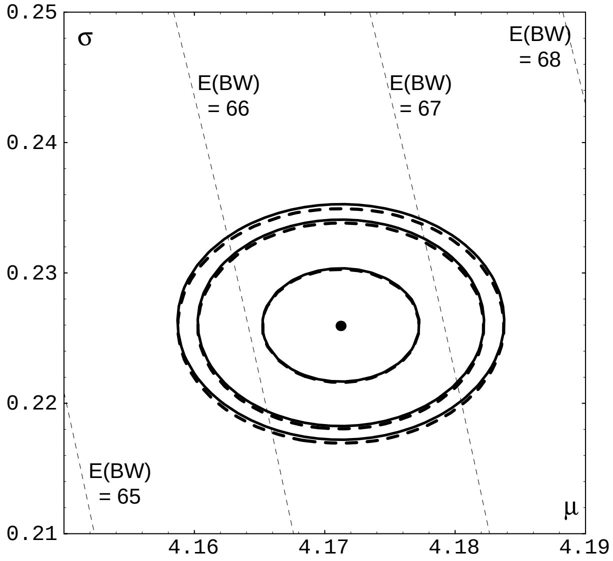

Figure 3.2-O: Joint Confidence Regions, n = 1,958

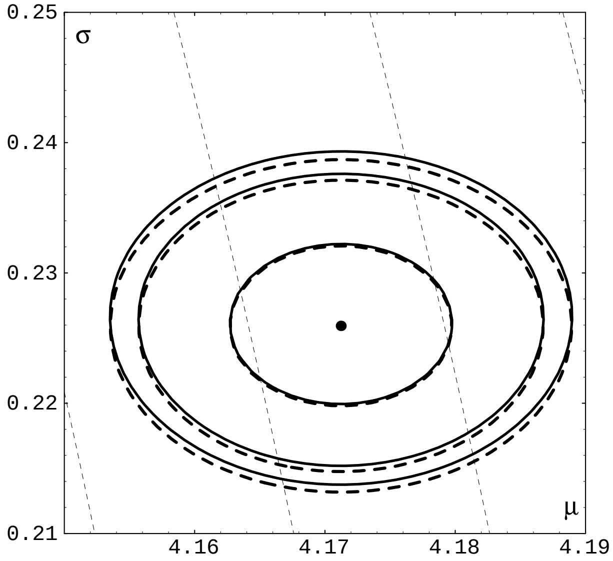

Figure 3.2-P: Joint Confidence Regions, n = 979

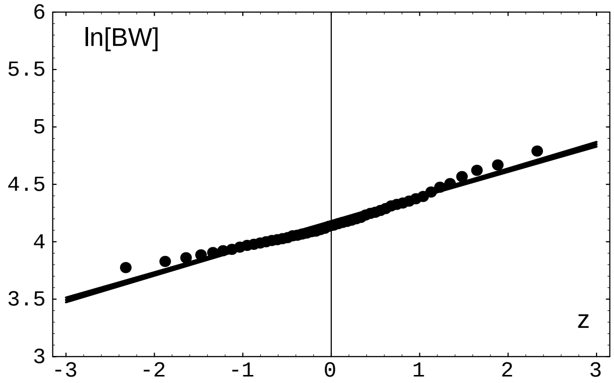

Figure 3.3-O: Confidence Bands, n = 1,958

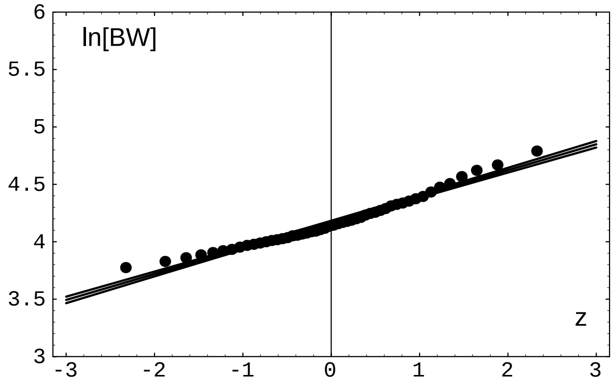

Figure 3.3-P: Confidence Bands, n = 979

Table 1
Data Set 1

| | Data Set 1 | Data Set 1 | Data Set 1 | Data Set 1 |
|---|---|---|---|---|
| | Original | Perturbed | Original | Perturbed |
| index | xi | x'i | ln(xi) | ln(x'i) |
| ••••• | ••••• | ••••• | ••••• | ••••• |
| 1 | 2 | 2 | 0.69 | 0.69 |
| 2 | 2 | 2 | 0.69 | 0.69 |
| 3 | 3 | 3 | 1.10 | 1.10 |
| 4 | 3 | 3 | 1.10 | 1.10 |
| 5 | 4 | 4 | 1.39 | 1.39 |
| 6 | 4 | 4 | 1.39 | 1.39 |
| 7 | 4 | 4 | 1.39 | 1.39 |
| 8 | 4 | 4 | 1.39 | 1.39 |
| 9 | 6 | 6 | 1.79 | 1.79 |
| 10 | 6 | 6 | 1.79 | 1.79 |
| 11 | 7 | 7 | 1.95 | 1.95 |
| 12 | 8 | 8 | 2.08 | 2.08 |
| 13 | 8 | 8 | 2.08 | 2.08 |
| 14 | 11 | 11 | 2.40 | 2.40 |
| 15 | 15 | 15 | 2.71 | 2.71 |
| 16 | 23 | 23 | 3.14 | 3.14 |
| 17 | 23 | 23 | 3.14 | 3.14 |
| 18 | 32 | 32 | 3.47 | 3.47 |
| 19 | 101 | 301 | 4.62 | 5.71 |
| ••••• | ••••• | ••••• | ••••• | ••••• |
| AMean | 14.00 | 24.53 | 2.01 | 2.07 |
| AStdDev | 22.66 | 67.47 | 1.02 | 1.19 |

exp[N[2,1]]

Table 2
Data Set 2

|  | Data Set 2 | Data Set 2 |
|---|---|---|
|  | Original | Perturbed |
| index | xi | x'i |
| ••••• | ••••• | ••••• |
| 1 | 0.04 | 0.04 |
| 2 | 0.06 | 0.06 |
| 3 | 0.09 | 0.09 |
| 4 | 0.10 | 0.10 |
| 5 | 0.12 | 0.12 |
| 6 | 0.14 | 0.14 |
| 7 | 0.15 | 0.15 |
| 8 | 0.18 | 0.18 |
| 9 | 0.20 | 0.20 |
| 10 | 0.21 | 0.21 |
| 11 | 0.22 | 0.22 |
| 12 | 0.26 | 0.26 |
| 13 | 0.29 | 0.29 |
| 14 | 0.30 | 0.30 |
| 15 | 0.31 | 0.31 |
| 16 | 0.32 | 0.32 |
| 17 | 0.32 | 0.32 |
| 18 | 0.33 | 0.33 |
| 19 | 0.38 | 0.38 |
| 20 | 0.40 | 0.40 |
| 21 | 0.41 | 0.41 |
| 22 | 0.42 | 0.42 |
| 23 | 0.52 | 0.52 |
| 24 | 0.57 | 0.57 |
| 25 | 0.61 | 0.75 |
| ••••• | ••••• | ••••• |
| AMean | 0.28 | 0.28 |
| AStdDev | 0.16 | 0.17 |

Beta[2,6]

Table 3
Data Set 3

|  | Body |  | Body |
|---|---|---|---|
| Percentile | Weight | Percentile | Weight |
|  | (kg) |  | (kg) |
| ••••• | ••••• | ••••• | ••••• |
| 0.01 | 43.6 | 0.51 | 63.1 |
| 0.03 | 46.0 | 0.53 | 63.9 |
| 0.05 | 47.5 | 0.55 | 64.6 |
| 0.07 | 48.8 | 0.57 | 65.2 |
| 0.09 | 49.7 | 0.59 | 65.9 |
| 0.11 | 50.5 | 0.61 | 66.7 |
| 0.13 | 51.1 | 0.63 | 67.5 |
| 0.15 | 52.1 | 0.65 | 69.0 |
| 0.17 | 52.9 | 0.67 | 69.9 |
| 0.19 | 53.4 | 0.69 | 70.6 |
| 0.21 | 54.1 | 0.71 | 71.7 |
| 0.23 | 54.6 | 0.73 | 72.9 |
| 0.25 | 55.2 | 0.75 | 74.6 |
| 0.27 | 55.7 | 0.77 | 75.6 |
| 0.29 | 56.1 | 0.79 | 76.6 |
| 0.31 | 56.6 | 0.81 | 77.8 |
| 0.33 | 57.5 | 0.83 | 79.4 |
| 0.35 | 57.9 | 0.85 | 81.1 |
| 0.37 | 58.5 | 0.87 | 84.2 |
| 0.39 | 58.9 | 0.89 | 87.9 |
| 0.41 | 59.6 | 0.91 | 90.6 |
| 0.43 | 59.9 | 0.93 | 96.4 |
| 0.45 | 60.7 | 0.95 | 101.7 |
| 0.47 | 61.4 | 0.97 | 106.7 |
| 0.49 | 62.5 | 0.99 | 120.3 |
|  |  | No = | 1,958 |
|  |  | Np = | 979 |