DOE Review Release 05/29/2025

Discussion of Variability in Health Physics Data

ORAUT-RPRT-0110 Rev. 00
Effective Date: 05/22/2025
Supersedes: None

Subject Expert(s): Nancy Chalmers, Thomas R. LaBone, and Daniel M. Sheanshang

Document Owner Approval:	Signature on File Nancy Chalmers, Document Owner	Approval Date:	05/09/2025
Concurrence:	Signature on File Wade C. Morris, Objective 1 Manager	Concurrence Date:	05/08/2025
Concurrence:	Signature on File Scott R. Siebert, Objective 3 Manager	Concurrence Date:	05/12/2025
Concurrence:	Signature on File Jennifer L. Hoff, Project Director	Concurrence Date:	05/08/2025
Approval:	Signature on File Brant Ulsh, Acting Associate Director of Science	Approval Date:	05/22/2025

FOR DOCUMENTS MARKED AS A TOTAL REWRITE OR REVISION,
REPLACE THE PRIOR REVISION AND DISCARD / DESTROY ALL COPIES OF THE PRIOR REVISION.

☒ New ☐ Total Rewrite ☐ Revision

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
05/22/2025	00	New draft report to discuss various methods to describe and quantify variability in data. Incorporates formal internal and NIOSH review comments. Training is not required. Initiated by Wade C. Morris and authored by Nancy Chalmers.

TABLE OF CONTENTS

SECTION **TITLE** **PAGE**

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

DOE	U.S. Department of Energy
GM	geometric mean
GSD	geometric standard deviation
IQR	interquartile range
LANL	Los Alamos National Laboratory
lm	linear model
mrem	millirem
NIOSH	National Institute for Occupational Safety and Health
ORAU	Oak Ridge Associated Universities
ORAUT	ORAU Team
PDF	probability density function
rlm	robust linear model
RSD	relative standard deviation
SRDB Ref ID	Site Research Database Reference Identification (number)

1.0 INTRODUCTION

In statistics, the population is the entire group of interest. Oftentimes, populations are large, and it is not feasible to collect data from each element of the population (census). A sample is a subset of the population from which data are actually collected. A parameter is a measure (usually unknowable) that describes the whole population, whereas a statistic is a measure (that can be calculated from collected data) that describes the sample. The goal of inference is to use the calculated sample statistic to estimate the unknown population parameter.

If elements of the population are heterogeneous, that inherent heterogeneity is referred to as "variability." If there is variability in the population, there will be variability in the sample because it is drawn from the population. If every element of the population is exactly the same with respect to the measure of interest (completely homogeneous), the population has no variability and no inference is necessary.

The "uncertainty" of an estimator is a measure of how much doubt one has about how that estimator will generalize to the entire population (i.e., how much doubt one has about how well the sample statistic estimates the population parameter).

As a very simple example, suppose the parameter of interest is the average height of adults in the United States (the population). The United States population is fairly large, and it is not feasible to measure the height of every adult (or have them self-report their height). The heights of adult Americans are heterogeneous, so there is variability in the population, and there will be variability in the sample. Measuring the heights of a sample of Americans would allow us to calculate a sample average (the statistic) to try to estimate the unknowable population average (the parameter). The uncertainty in the sample average (the estimator) takes into account things like the sample size, how the sample was chosen, and the variability in the sample, amongst other things. For example, all other things being equal, a very small sample would have more uncertainty (cast more doubt about how well the sample average estimates the population average) than a very large sample.

On this Project, it is common to assume that a population of data is lognormally distributed. For lognormal distributions, the most common measure of variability (or spread in the data) is the geometric standard deviation (GSD). Since we are typically dealing with a sample of data from the population, the GSD of a lognormal fit is a sample statistic (an estimator with its own uncertainty) that attempts to describe the GSD of the population (parameter) from which the data were sampled. Similarly, the geometric mean (GM) of a lognormal fit is a sample statistic (an estimator with its own uncertainty) that attempts to describe the GM of the population (parameter) from which the data were sampled. The GSD (whether parameter or statistic) is NOT the uncertainty in the GM.

When reporting sample GSDs, one question might be, "how big a GSD is too big?" GSD is a measure of variability in a set of data. Variability is an inherent characteristic of the data. This Project deals with historical data (already collected), meaning there is no opportunity to control the process that generated the data, so reducing variability is not possible[1]. It is, however, important to adequately characterize and quantify the variability present in these historical datasets.

The purpose of this report is to explore ways to characterize and quantify variability. Section 2.0 presents several measures of variability. Section 3.0 discusses some possible causes of large variability. Section 4.0 presents the lognormal distribution and discusses its use in health physics. Sections 5.0 and 6.0 describe mixture distributions that can be useful when fitting health physics data. Section 7.0 introduces the examples. Sections 8.0, 9.0, and 10.0 include analyses of three health

---

[1] Historically, operational health physicists could have attempted to control variability in exposure data by using exposure controls (respiratory protection, shielding, etc.)

physics datasets (bioassay data, dosimeter data, and environmental data, respectively). Section 11.0 presents some conclusions. All simulations and analyses are further documented in the support files for this report [Oak Ridge Associated Universities (ORAU) Team (ORAUT) 2024].

2.0 SUMMARY STATISTICS

Summary statistics tend to fall into three categories:

1. Counts or tallies,
2. Measures of central tendency, and
3. Measures of variability (also called dispersion).

While the focus of this report is variability, measures of variability describe the spread of data around some central value, so measures of central tendency must also be discussed. For all the following summary statistics, n is the sample size or number of observations. The observations (or data) are represented as $x_1, x_2, \ldots, x_n$. All summary statistics can be found in Ellison et al. [2009, pp. 40–44], unless otherwise referenced.

2.1 MEASURES OF CENTRAL TENDENCY

2.1.1 Arithmetic Mean

The arithmetic mean is also known as simply the mean or the average of a set of data. The arithmetic mean is:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2-1}$$

2.1.2 Geometric Mean

In the context of summary statistics, the GM has nothing to do with a lognormal distribution. A sample GM can be calculated for any positive set of data [Gilbert 1987, pp. 167, 185]. The sample GM is:

$$GM = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(x_i\right)\right] = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{2-2}$$

The middle expression is just the exponentiated arithmetic mean of the logarithms[2] of the data. The expression on the right is another formulation of the GM that may also be familiar.

2.1.3 Median

In general, a quantile Q_p is a value wherein at least a fraction p of the data is less than or equal to the quantile *and* at least $(1 - p)$ of the data is greater than or equal to the quantile, where $0 < p < 1$. For example, the median (also known as the 50th percentile or the number that divides the dataset in half) is $Q_{0.5}$. That means that at least 0.5 (or half) of the data is less than or equal to the median *and* at least 0.5 (half) of the data is greater than or equal to the median.

---

[2] All logarithms in this report are natural logarithms.

There are nine methods to calculate sample quantiles [National Institute for Occupational Safety and Health (NIOSH) 2019, p. 9]. This report uses the Type 7 method, which is the default in the R software.

2.2 MEASURES OF VARIABILITY

2.2.1 Variance

Variance is a measure of the spread of observations around the arithmetic mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \qquad (2\text{-}3)$$

2.2.2 Standard Deviation

Standard deviation is another measure of the spread of observations around the arithmetic mean. It is the positive square root of the variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \qquad (2\text{-}4)$$

The standard deviation is usually more useful to report than variance because standard deviation has the same units as the observed data, whereas variance has squared units.

2.2.3 Relative Standard Deviation (or Coefficient of Variation)

The relative standard deviation (RSD; or coefficient of variation) is a measure of the spread of data in comparison to the arithmetic mean:

$$RSD = CV = \frac{s}{\overline{x}} \qquad (2\text{-}5)$$

The RSD is formulated here as a fraction, but it is often expressed as a percentage.

2.2.4 Geometric Standard Deviation

In the context of summary statistics, the GSD has nothing to do with a lognormal distribution. A sample GSD can be calculated for any positive set of data [Gilbert 1987, p. 167]. The sample GSD is:

$$GSD = \exp\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(x_i\right) - \log\left(GM\right)\right)^2}\right] \qquad (2\text{-}6)$$

This is very similar to the standard deviation formula (Equation 2-4), except that GSD is the exponentiation using the logarithms of the data and the logarithm of the GM as the central tendency. Also note that the GSD has *n* in the denominator, while the standard deviation has *n*-1.

2.2.5 Range

The range is the difference between the largest and smallest observed values:

$$range = maximum - minimum \qquad (2\text{-}7)$$

Range is sometimes expressed as an interval:

$$(minimum, maximum) \tag{2-8}$$

2.2.6 Interquartile Range

Quartiles divide data into quarters. The three quartiles are $Q_{0.25}$, $Q_{0.5}$, and $Q_{0.75}$ (the 25th, 50th, and 75th percentiles, respectively); $Q_{0.5}$ is the median, $Q_{0.25}$ and $Q_{0.75}$ are the limits of the box portion of a boxplot, and $Q_{0.5}$ is the line inside the box. The interquartile range (IQR) is:

$$IQR = Q_{0.75} - Q_{0.25} \tag{2-9}$$

2.2.7 Ratio of Quantiles

Ratios of quantiles are used in biomedical applications, cost-effectiveness analysis, environmental applications, ceramic materials studies, and economics (e.g. $Q_{0.8}/Q_{0.2}$ or $Q_{0.9}/Q_{0.4}$ to quantify variability in income inequality) [Farcomeni and Geraci 2024, p. 2]. This report uses $Q_{0.95}/Q_{0.05}$, the ratio of the 95th percentile to the 5th percentile, to describe the spread of most of the data. If the 5th percentile is negative, zero, or censored, that ratio is not as useful as a positive uncensored ratio. The 50th percentile is less likely to be negative, zero, or censored, so $Q_{0.95}/Q_{0.5}$, the ratio of the 95th percentile to the 50th percentile, is also used to describe the spread of most of the larger half of the data.

3.0 CAUSES OF LARGE VARIABILITY

With historical data, reducing variability is impossible, but it is important to adequately characterize and quantify the variability. The typical concern is with what is perceived as large variability in data or a fitted distribution (e.g., a large GSD of a lognormal fit). Three possible causes of large variability are addressed in this section:

1. Outliers,
2. Need to stratify, and
3. Wrong fitted distribution.

3.1 OUTLIERS

An outlier is an observation in a set of data that appears to be inconsistent with the remainder of that dataset. Outliers can be caused by human error in measurement, instrument failure, or typographical error. There is nothing inherently bad about outliers, though they should always be checked for transcription errors and relevance. Although there are several outlier tests available in the literature and software, rather than test and expunge outliers, a robust analysis is less affected by outliers [Ellison et al. 2009, pp. 68–69]. Random variation occasionally produces extreme values by chance. If a perceived outlier is valid, the variance will be underestimated if the outlier is expunged [Venables and Ripley 2002, pp. 130–131].

Robust sample statistics are designed to be less sensitive to outliers than their nonrobust counterparts. For example, the arithmetic mean can be upset by a single outlier. The median will tolerate up to 50% outliers. The median is a more robust measure of central tendency than the arithmetic mean. The variance and standard deviation are similarly less robust than measures of variability that are based on quantiles (IQR and ratios of quantiles) [Venables and Ripley 2002, pp. 130–131].

Another useful robust method is robust regression. In regression, an outlier is sometimes referred to as an influence point, if by virtue of its position it has an overriding influence on the fit. Instead of the typical regression methods, robust regression methods can down-weight questionable observations, so there is no need to decide whether to expunge them. The “rlm” function[3] in R uses an iterated reweighted least squares method that uses starting weights (or defaults) and iterates until it finds a suitable solution [Huber 1981, pp. 172–174].

The usefulness of robust regression can be seen through simulation. Suppose there are 30 sets of paired data that have a fairly strong linear relationship (see Figure 3-1). The 31st observation appears to be inconsistent with the remainder of the data. Ideally, the regression line would reflect the fairly strong linear relationship of the majority of the points. In Figure 3-2, the blue dashed line and the first set of parameters (labeled “lm”[4]) are from a typical linear regression. Note that the blue dashed line is pulled away from the relationship of the majority of the points because of the one inconsistent result that has influence on the fit. In Figure 3-2, the red solid line and the second set of parameters (labeled “rlm”) are from a robust linear regression. The one inconsistent result is down-weighted, so the red solid line does a good job representing the relationship of the majority of the data. An alternative to robust regression is quantile regression [ORAUT 2018, pp. 23–27]. In Figure 3-2, the purple dot-dashed line and the third set of parameters (labeled “50th”) are from a quantile regression, which is median regression. That line and parameters are similar to the robust regression line and parameters. The choice between robust regression and quantile regression is mostly determined by how the fit will be used. For the remainder of this report, all regression is robust regression, so perceived outliers have less influence on the fits.

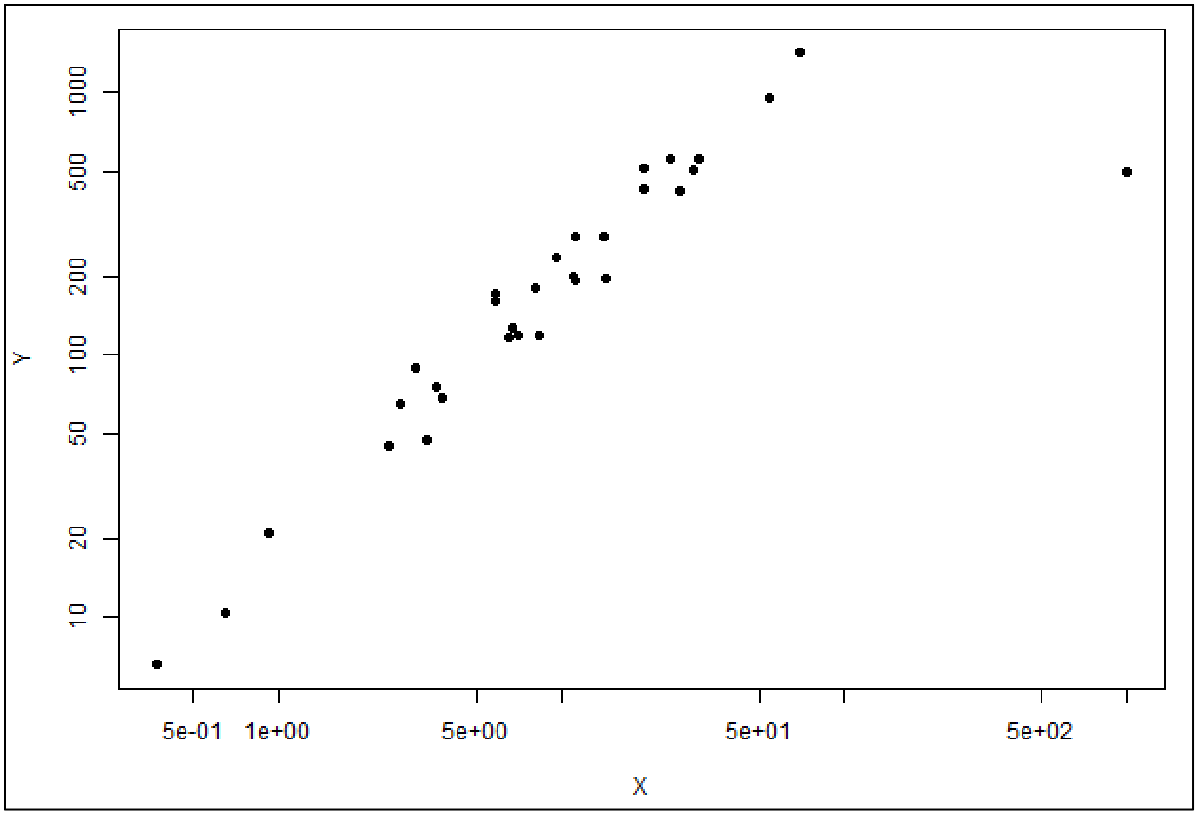

Figure 3-1. Scatterplot of 31 simulated paired results. Attachment A contains an extended description.

---

[3] rlm stands for robust linear model.

[4] lm stands for linear model.

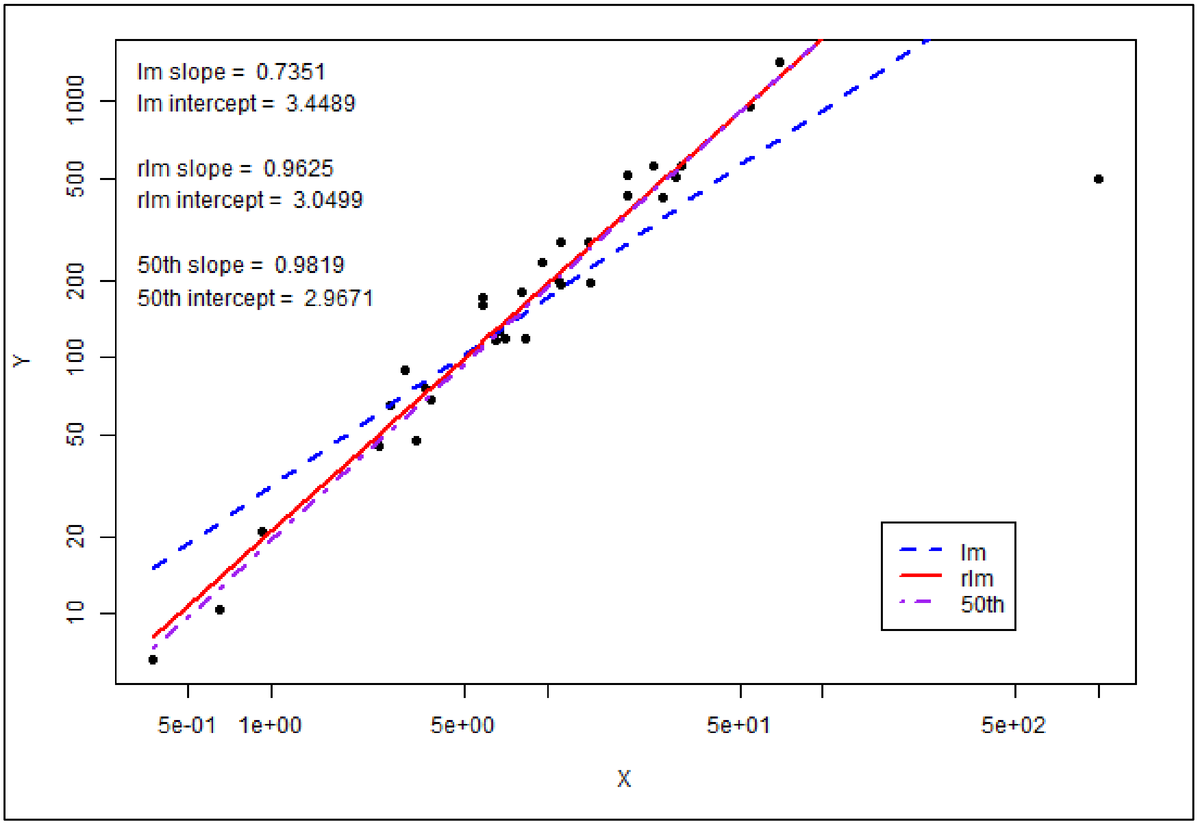

Figure 3-2. Scatterplot of 31 simulated paired results with parameters and lines fit using regression, robust regression, and quantile (median) regression. Attachment A contains an extended description.

3.2 NEED TO STRATIFY

If a dataset consists of multiple distinct homogeneous groups (little variability within a group but differences between the groups), the combined dataset might appear to have a large amount of variability. These relatively homogeneous subgroups are called strata. Stratification aims to break a heterogeneous dataset into multiple relatively homogenous strata. The strata must be mutually exclusive and exhaustive. If each stratum is homogenous, the variability of each stratum will be smaller than the variability of the whole dataset.

Stratification decisions should never be made solely by looking at the quantitative results to be analyzed. In other words, the analyst should not stratify the highest half of the quantitative results from the lowest half based only on the quantitative results. Stratification should be based on what the health physicist knows about the site, processes, and data that would suggest that there are distinct homogeneous subgroups. The dataset might contain categorical information (e.g., sample type, worker type, detector type) that would facilitate meaningful stratification. Sometimes that information might not be available in the dataset or in any ancillary source. A stratification decision should not be based on a data-mining mission by the analyst. If a dataset is stratified into unmeaningful subgroups, the resulting analysis will be less precise than using all the data because of the smaller sample sizes [ORAUT 2014, p. 10].

Stratification cannot be considered in the simulated examples in Sections 4.0, 5.0, and 6.0. Each of the health physics dataset examples in Sections 8.0, 9.0, and 10.0 explore stratification by whatever categorical variables are available. For the dataset examples, no comment is made on the appropriateness of stratification as an analysis method; it is included as an option for illustrative purposes.

3.3 WRONG FITTED DISTRIBUTION

Fitting the right probability distribution to a dataset is important because it is the basis for inferences and predictions. If a probability distribution does not align well with the characteristics of the data, the fitted parameters of the distribution do a poor job describing the data. For example, if data are not lognormally distributed, but a lognormal distribution is fit to them anyway, the fitted lognormal parameters (GM and GSD) might not describe the data well. In fact, as shown in subsequent sections, the variability described by the GSD of the lognormal fit can appear to be large, but the data were not lognormally distributed to begin with. The normal-lognormal mixture and lognormal-lognormal mixture distributions are presented as alternatives to the lognormal distribution.

4.0 LOGNORMAL DISTRIBUTION

The lognormal distribution is widely used in health physics. It is commonly used when data are positive and skewed. If data are lognormally distributed, the logarithms of those data are normally distributed. The probability density function (PDF) for a lognormal distribution is:

$$f_{\mathrm{LN}}(x) = \frac{1}{x\sqrt{2\pi\sigma_{\mathrm{LN}}^2}} \exp\left[-\frac{1}{2\sigma_{\mathrm{LN}}^2}\left(\log\ (x) - \mu_{\mathrm{LN}}\right)^2\right] \tag{4-1}$$

where

$0 < x < \infty$
$-\infty < \mu_{\mathrm{LN}} < \infty$
$\sigma_{\mathrm{LN}} > 0$
$\mathrm{GM} = \exp(\mu_{\mathrm{LN}}) > 0$
$\mathrm{GSD} = \exp(\sigma_{\mathrm{LN}}) > 1$

Some notes on the lognormal distribution:

- Throughout this report, the lognormal distribution is parameterized using the GM and GSD, instead of the less familiar “log mean” (μ_{LN}) and “log standard deviation” (σ_{LN}).
- If data are perfectly lognormally distributed, the fitted GM and fitted GSD will equal the sample GM and sample GSD from Equations 2-2 and 2-6, respectively.
- The fitted GSD is a measure of the variability (or spread) of the fitted distribution and is *not* the uncertainty in the fitted GM (see discussion in Section 1.0). The fitted GM and fitted GSD both have their own uncertainty.
- The lognormal distribution is used across various branches of the sciences. A study of GSDs across the sciences summarizes GSD values ranging from 1.1 to 33, with most of the GSDs above 3 from species abundance studies from ecology [Limpert et al. 2001, pp. 7–9].

Simulation
Although the lognormal distribution is familiar, this report presents a simple simulation to introduce plots and concepts for some less familiar distributions in subsequent sections. One thousand results were simulated from a lognormal distribution with GM = 5 and GSD = 3. Figure 4-1 is a histogram summarizing those 1,000 results, with the lognormal PDF (Equation 4-1) overlaid as the red density curve.

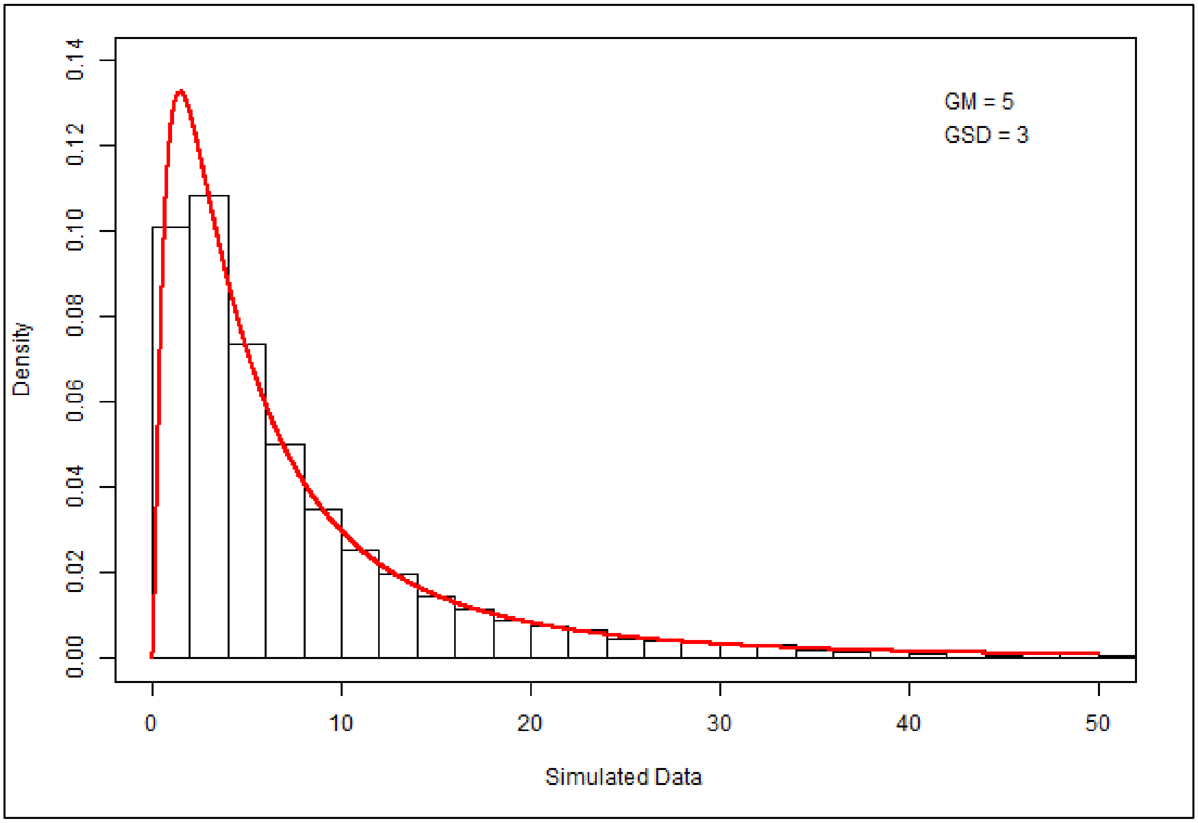

Figure 4-1. Histogram of 1,000 simulated results with true lognormal PDF overlaid. Attachment A contains an extended description.

A lognormal distribution was fit to the 1,000 simulated results. Figure 4-2 shows the same histogram from Figure 4-1, but the overlaid curve in Figure 4-2 is the PDF from the lognormal fit. Figure 4-3 is a lognormal probability plot (or quantile-quantile plot) that displays the same information as Figure 4-2. Figure 4-3 allows us to judge the fit of the distribution by comparing the data to a line, rather than comparing a binned histogram to a curve as in Figure 4-2. Unsurprisingly, based on both Figures 4-2 and 4-3, the lognormal distribution fits well and the fitted parameters closely match the known parameters for the simulated data.

Table 4-1 summarizes various measures of variability of the true distribution, the simulated data, and the fitted distribution. As expected, the three columns have good agreement for all the measures of variability.

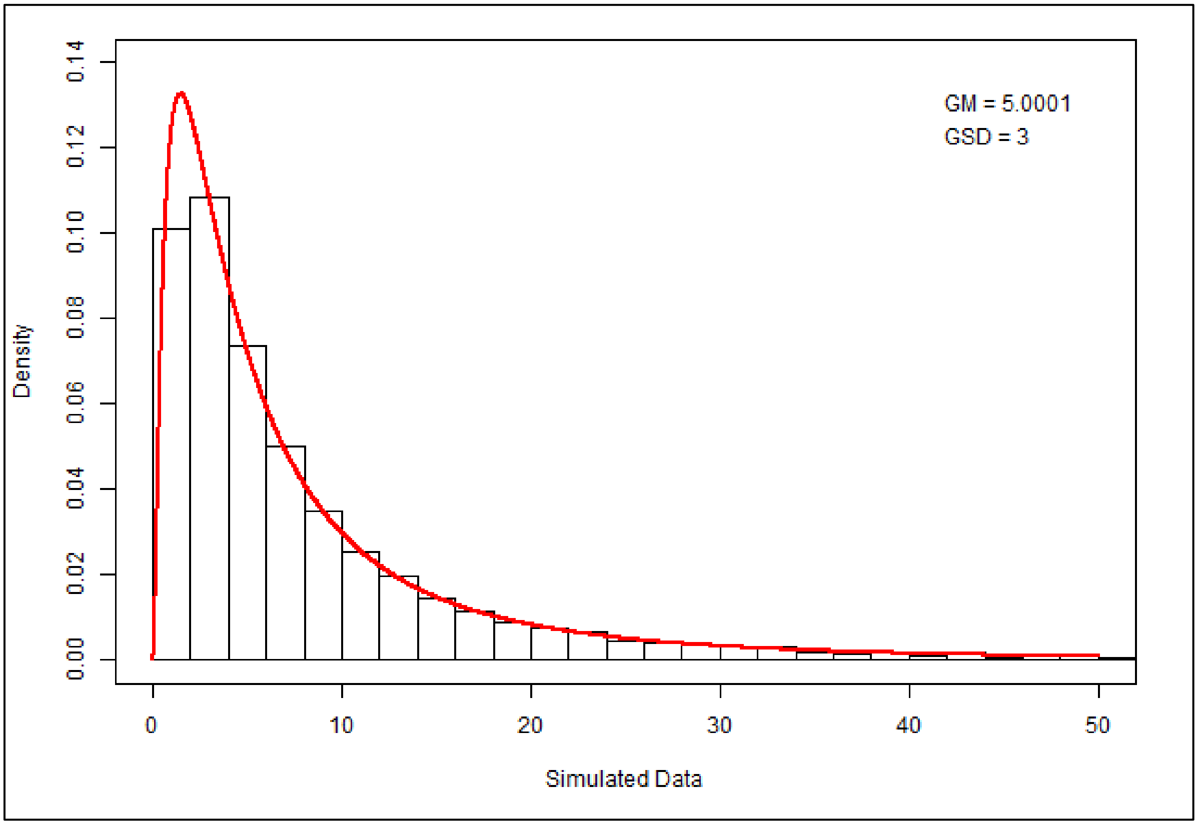

Figure 4-2. Histogram of 1,000 simulated results with fitted lognormal PDF overlaid. Attachment A contains an extended description.

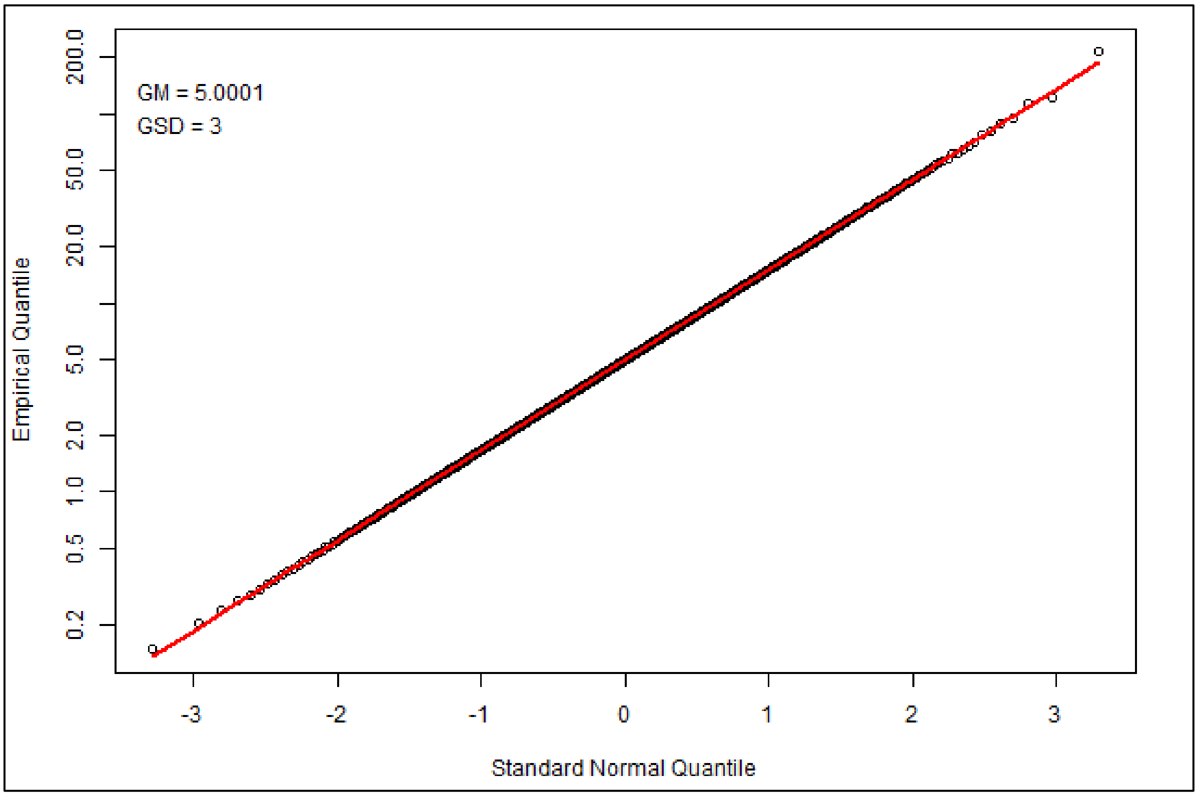

Figure 4-3. Lognormal probability plot of 1,000 simulated results with fitted lognormal. Attachment A contains an extended description.

Table 4-1. Summary of various measures of variability for lognormal distribution.

Measure of variability	True distribution	Simulated data (empirical)	Fitted lognormal
Range	(0, ∞)	(0.1474, 212.3982)	(0, ∞)
IQR	8.1070	8.0971	8.1071
Variance	195.8544	186.3938	195.8593
Standard deviation	13.9948	13.6526	13.9950
$Q_{0.95}/Q_{0.05}$	37.1184	36.6217	37.1182
$Q_{0.95}/Q_{0.5}$	6.0925	6.053	6.0925

5.0 NORMAL-LOGNORMAL MIXTURE DISTRIBUTION

The normal-lognormal mixture distribution is a mixture of a normal distribution and a lognormal distribution. Sometimes, health physics data consist of results from workers who were not exposed (noise that can be described by a normal distribution) and results from workers who were exposed (signal that can be described by a lognormal distribution). Attempting to describe both noise and signal by only a normal distribution or only a lognormal distribution might be inadequate. The PDF for a normal-lognormal mixture distribution is:

$$f_{\mathrm{NLN}}(x) = a_{\mathrm{N}} f_{\mathrm{N}}(x) + (1 - a_{\mathrm{N}}) f_{\mathrm{LN}}(x) \qquad (5\text{-}1)$$

where

a_{N} = the mixing fraction, the fraction that is normal
$(1 - a_{\mathrm{N}})$ = the fraction that is lognormal

$$f_{\mathrm{N}}(x) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{N}}^2}} \exp\left[-\frac{1}{2\sigma_{\mathrm{N}}^2}(x - \mu_{\mathrm{N}})^2\right] \qquad (5\text{-}2)$$

where

$-\infty < x < \infty$
$\sigma_{\mathrm{N}} > 0$
$-\infty < \mu_{\mathrm{N}} < \infty$

$$f_{\mathrm{LN}}(x) = \frac{1}{x\sqrt{2\pi\sigma_{\mathrm{LN}}^2}} \exp\left[-\frac{1}{2\sigma_{\mathrm{LN}}^2}\left(\log(x) - \mu_{\mathrm{LN}}\right)^2\right] \qquad (5\text{-}3)$$

where

$0 < x < \infty$
$\sigma_{\mathrm{LN}} > 0$
$-\infty < \mu_{\mathrm{LN}} < \infty$
$\mathrm{GM} = \exp(\mu_{\mathrm{LN}}) > 0$
$\mathrm{GSD} = \exp(\sigma_{\mathrm{LN}}) > 1$

Simulation

One thousand results were simulated from a normal-lognormal mixture distribution with mixing fraction $a_{\mathrm{N}} = 0.5$, $\mu_{\mathrm{N}} = 0$, $\sigma_{\mathrm{N}} = 0.5$, GM = 5 and GSD = 3. Figure 5-1 is a histogram summarizing those 1,000 results, with the normal-lognormal mixture PDF (Equation 5-1) overlaid as the red density curve.

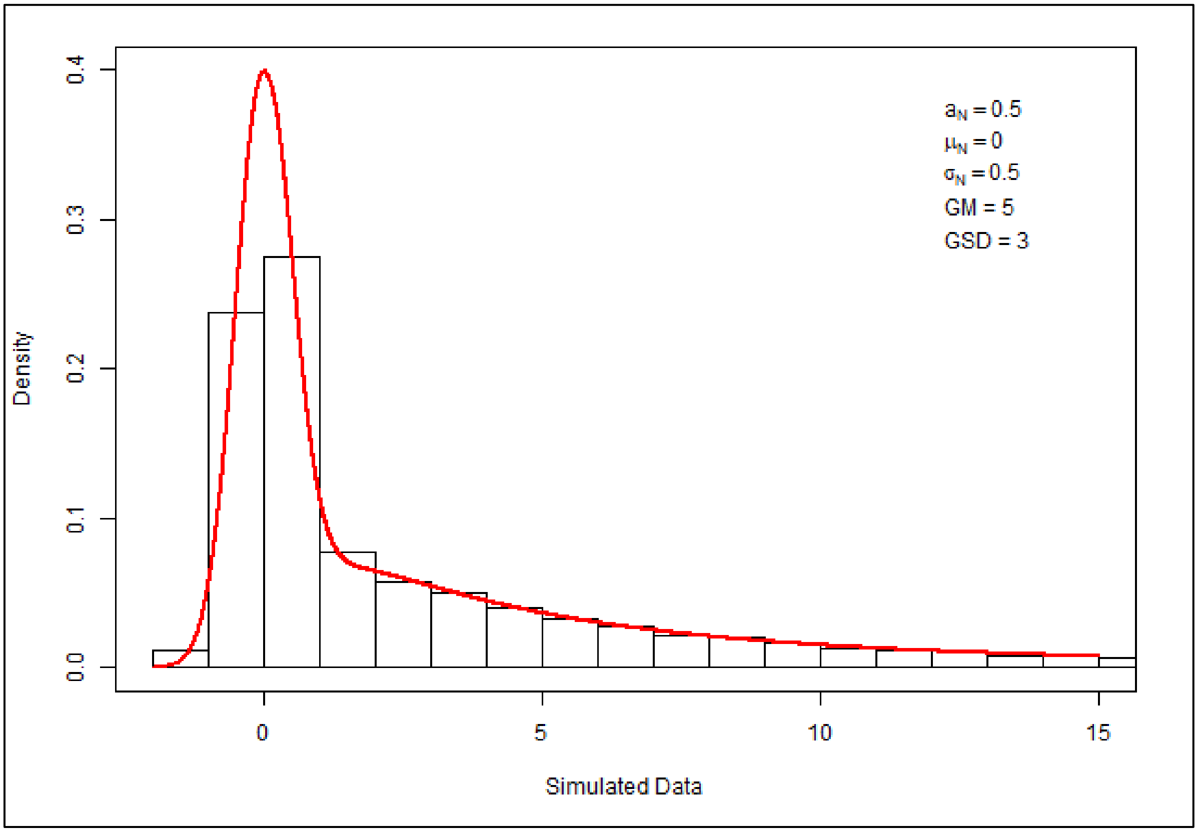

Figure 5-1. Histogram of 1,000 simulated results with true normal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

For illustrative purposes, a lognormal distribution was fit to the 1,000 simulated results. Figure 5-2 shows the same histogram from Figure 5-1, but the overlaid curve in Figure 5-2 is the PDF from a lognormal fit. Figure 5-3 is a lognormal probability plot. The data were simulated from a normal-lognormal mixture so, unsurprisingly and based on both Figures 5-2 and 5-3, the lognormal distribution does not fit well. The lognormal distribution has no way of handling the negative data that the normal-lognormal mixture can produce. Note that the GSD of the lognormal fit is 9.5515, while the GSD of the lognormal portion of the true normal-lognormal mixture is 3.

For illustrative purposes, a normal distribution was fit to the 1,000 simulated results. Figure 5-4 shows the same histogram from Figure 5-1, but the overlaid curve in Figure 5-4 is the PDF from a normal fit. Figure 5-5 is a normal probability plot. The data were simulated from a normal-lognormal mixture so, unsurprisingly and based on both Figures 5-4 and 5-5, the normal distribution does not fit well.

A normal-lognormal mixture distribution was fit to the 1,000 simulated results. Figure 5-6 shows the same histogram from Figure 5-1, but the overlaid curve in Figure 5-6 is the PDF from a normal-lognormal mixture fit. Figure 5-7 is a probability plot. Unsurprisingly, based on both Figures 5-6 and 5-7, the normal-lognormal mixture distribution fits well, and the fitted parameters closely match the known parameters for the simulated data.

Table 5-1 summarizes various measures of variability of the true distribution, the simulated data, and the fitted distributions. As one would expect, the normal-lognormal mixture column has much better agreement with the true and empirical columns than do the lognormal and normal columns.

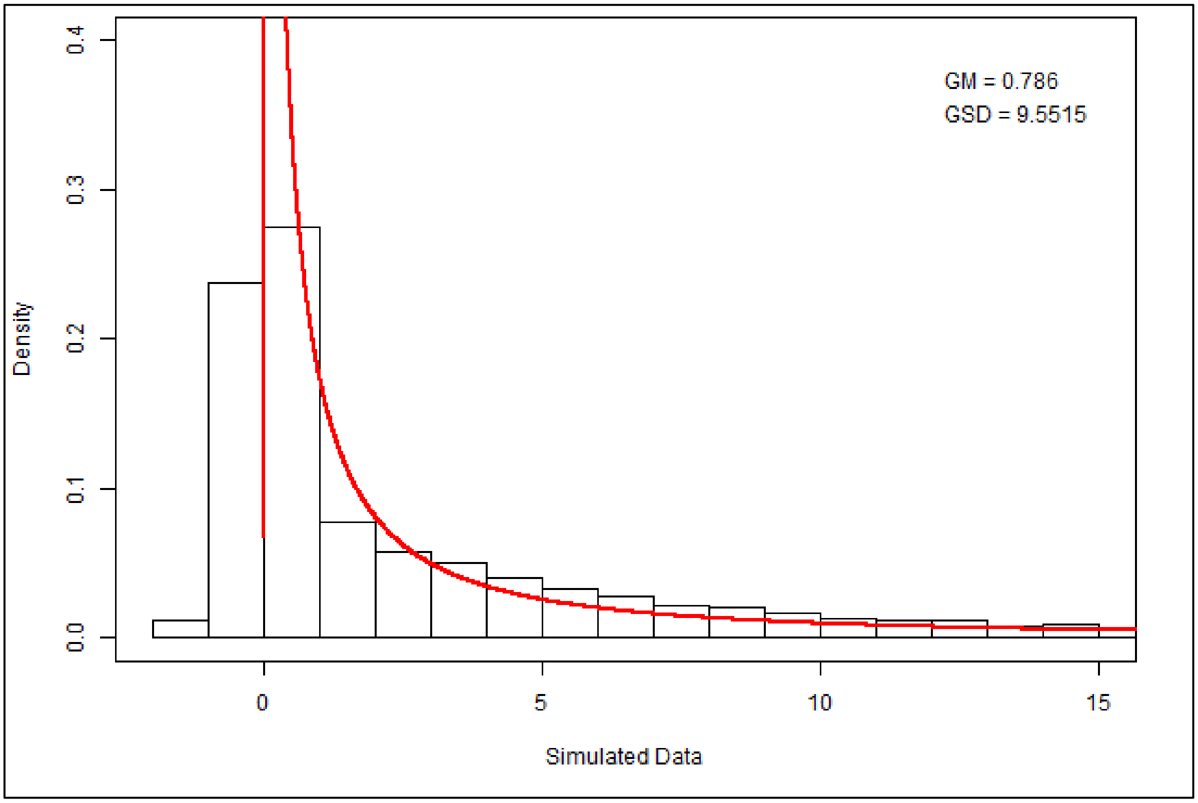

Figure 5-2. Histogram of 1,000 simulated results with fitted lognormal PDF overlaid. Attachment A contains an extended description.

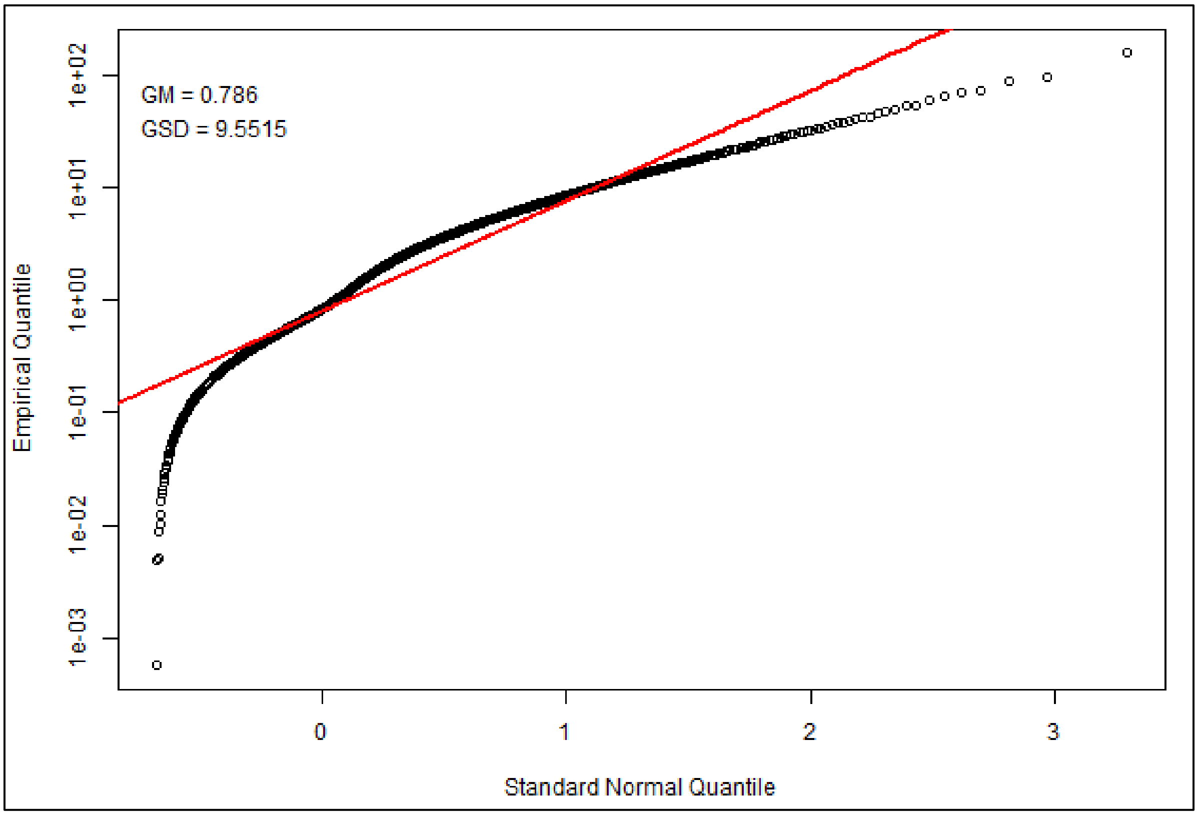

Figure 5-3. Lognormal probability plot of 1,000 simulated results with fitted lognormal. Attachment A contains an extended description.

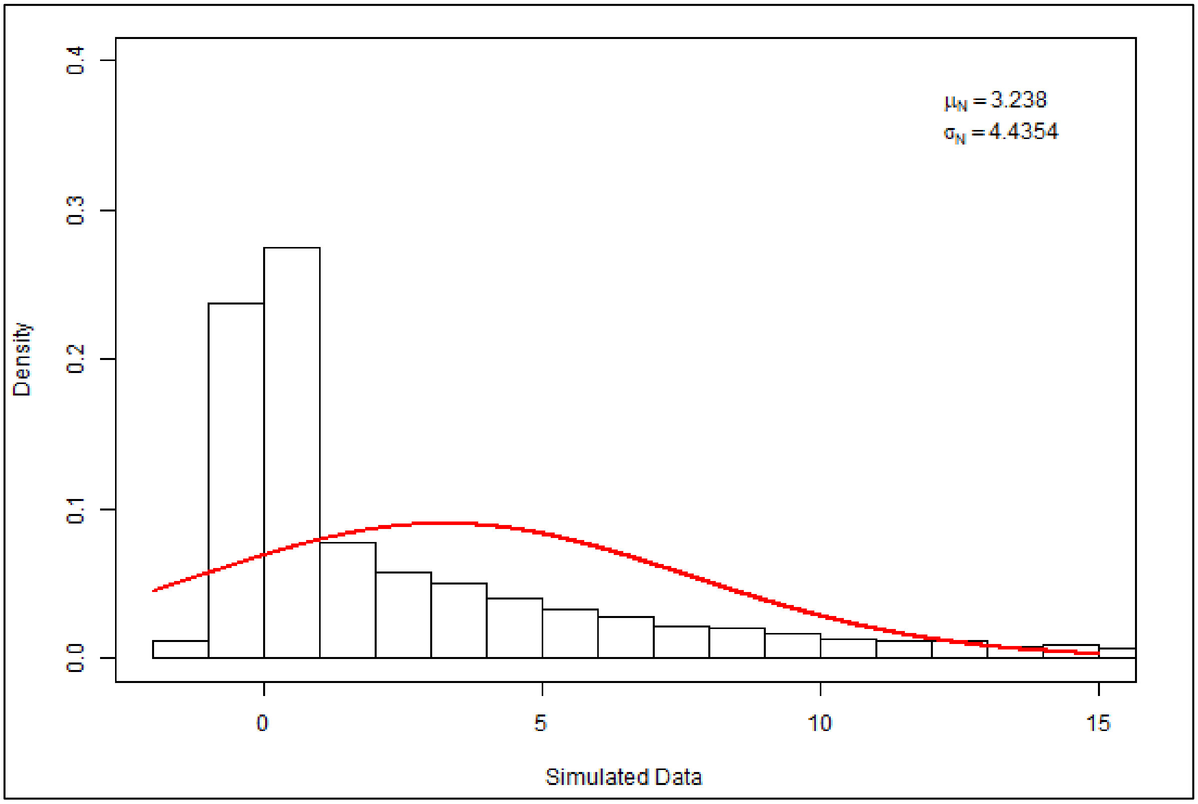

Figure 5-4. Histogram of 1,000 simulated results with fitted normal PDF overlaid. Attachment A contains an extended description.

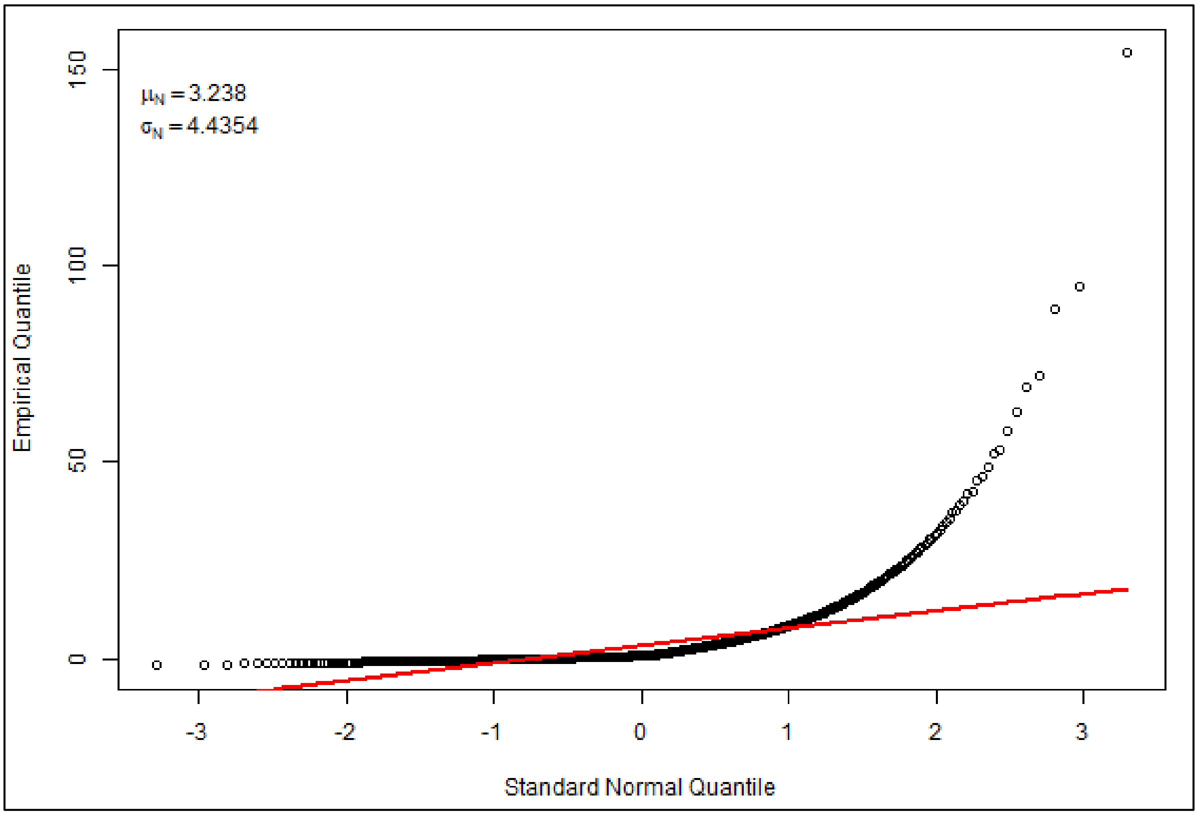

Figure 5-5. Normal probability plot of 1,000 simulated results with fitted normal. Attachment A contains an extended description.

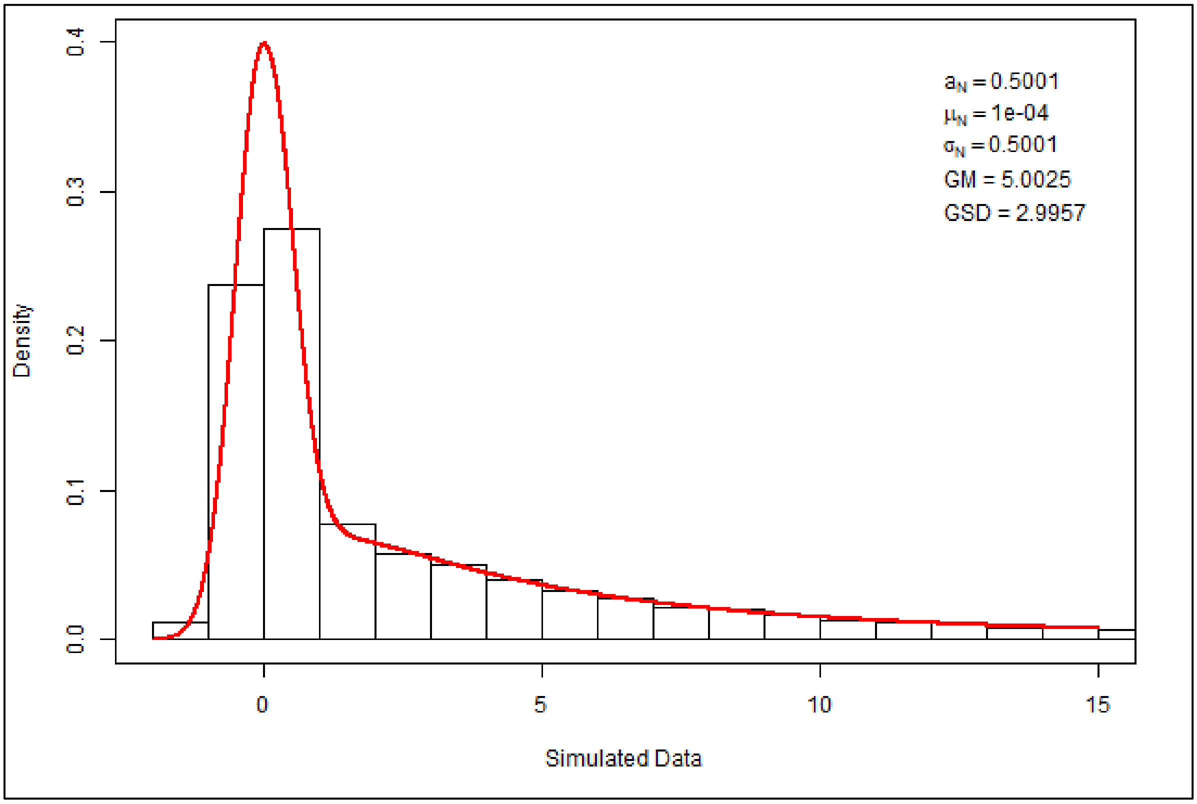

Figure 5-6. Histogram of 1,000 simulated results with fitted normal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

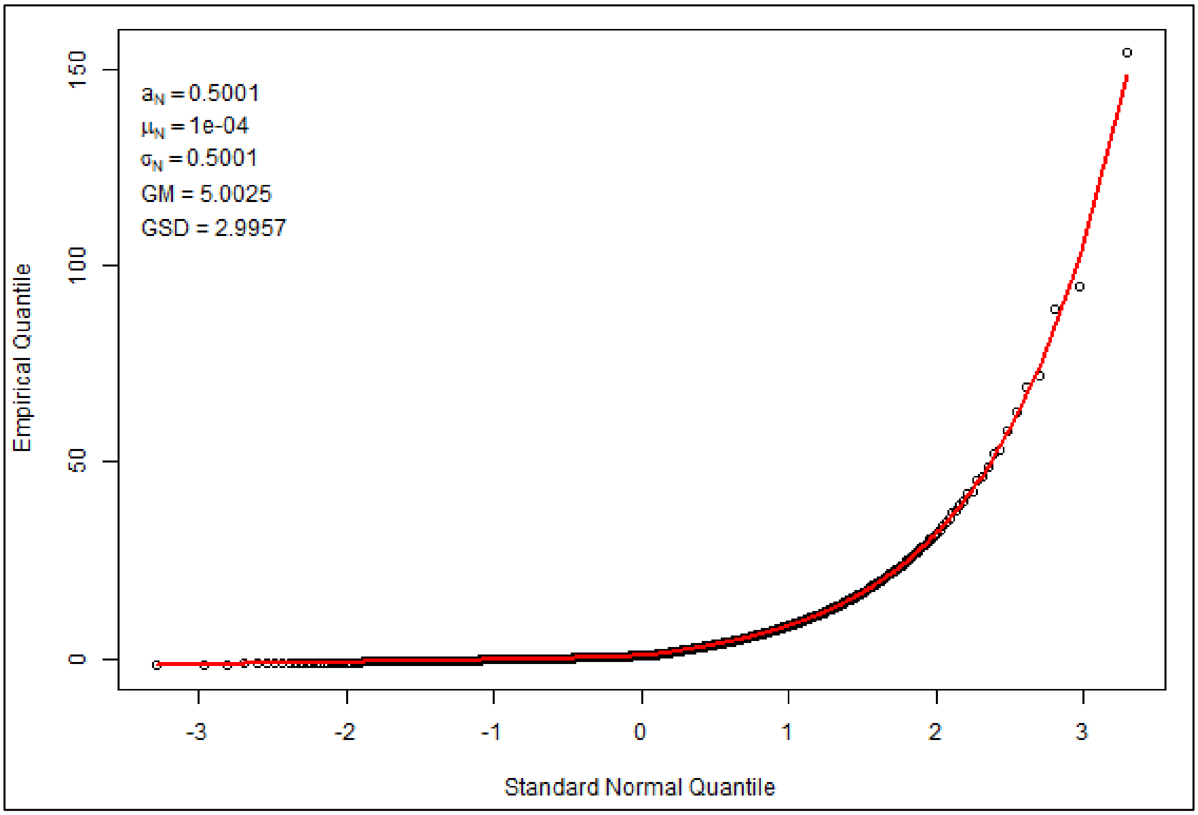

Figure 5-7. Probability plot of 1,000 simulated results with fitted normal-lognormal mixture. Attachment A contains an extended description.

Table 5-1. Summary of various measures of variability for normal-lognormal mixture distribution.

Measure of variability	True distribution	Simulated data (empirical)	Fitted lognormal	Fitted normal	Fitted normal-lognormal mixture
Range	(−∞, ∞)	(−1.6012, 153.9647)	(0, ∞)	(−∞, ∞)	(−∞, ∞)
IQR	5	4.9954	3.43	5.9833	5.0007
Variance	118.9476	107.1408	16,280.6256	19.6729	118.2256
Standard deviation	10.9063	10.3509	127.5956	4.4354	10.8732
$Q_{0.95}/Q_{0.05}$	−31.8946	−31.9952	1,675.5420	−2.5961	−31.8394
$Q_{0.95}/Q_{0.5}$	24.8680	24.7836	40.9334	3.2531	24.8227

6.0 LOGNORMAL-LOGNORMAL MIXTURE DISTRIBUTION

The lognormal-lognormal mixture distribution is a mixture of two lognormal distributions. Sometimes, health physics data consist of results from more than one lognormal distribution. Fitting a single lognormal distribution to a lognormal-lognormal mixture might be inadequate. The PDF for a lognormal-lognormal mixture distribution is:

$$f_{\text{LNLN}}(x) = a_1 f_{\text{LN1}}(x) + (1 - a_1) f_{\text{LN2}}(x) \tag{6-1}$$

where

a_1 is the mixing fraction, the fraction that is the first lognormal
$(1 - a_1)$ is the fraction that is the second lognormal

$$f_{\text{LN1}}(x) = \frac{1}{x\sqrt{2\pi\sigma_{\text{LN1}}^2}} \exp\left[-\frac{1}{2\sigma_{\text{LN1}}^2}\left(\log(x) - \mu_{\text{LN1}}\right)^2\right] \tag{6-2}$$

where

$0 < x < \infty$
$\sigma_{\text{LN1}} > 0$
$-\infty < \mu_{\text{LN1}} < \infty$
$\text{GM}_1 = \exp(\mu_{\text{LN1}}) > 0$
$\text{GSD}_1 = \exp(\sigma_{\text{LN1}}) > 1$

$$f_{\text{LN2}}(x) = \frac{1}{x\sqrt{2\pi\sigma_{\text{LN2}}^2}} \exp\left[-\frac{1}{2\sigma_{\text{LN2}}^2}\left(\log(x) - \mu_{\text{LN2}}\right)^2\right] \tag{6-3}$$

where

$0 < x < \infty$
$\sigma_{\text{LN2}} > 0$
$-\infty < \mu_{\text{LN2}} < \infty$
$\text{GM}_2 = \exp(\mu_{\text{LN2}}) > 0$
$\text{GSD}_2 = \exp(\sigma_{\text{LN2}}) > 1$

Simulation

One thousand results were simulated from a lognormal-lognormal mixture distribution with mixing fraction $a1 = 0.25$, $\text{GM}_1 = 2$, $\text{GSD}_1 = 2$, $\text{GM}_2 = 20$, and $\text{GSD}_2 = 2$. Figure 6-1 is a histogram summarizing those 1,000 results, with the lognormal-lognormal mixture PDF (Equation 6-1) overlaid as the red density curve.

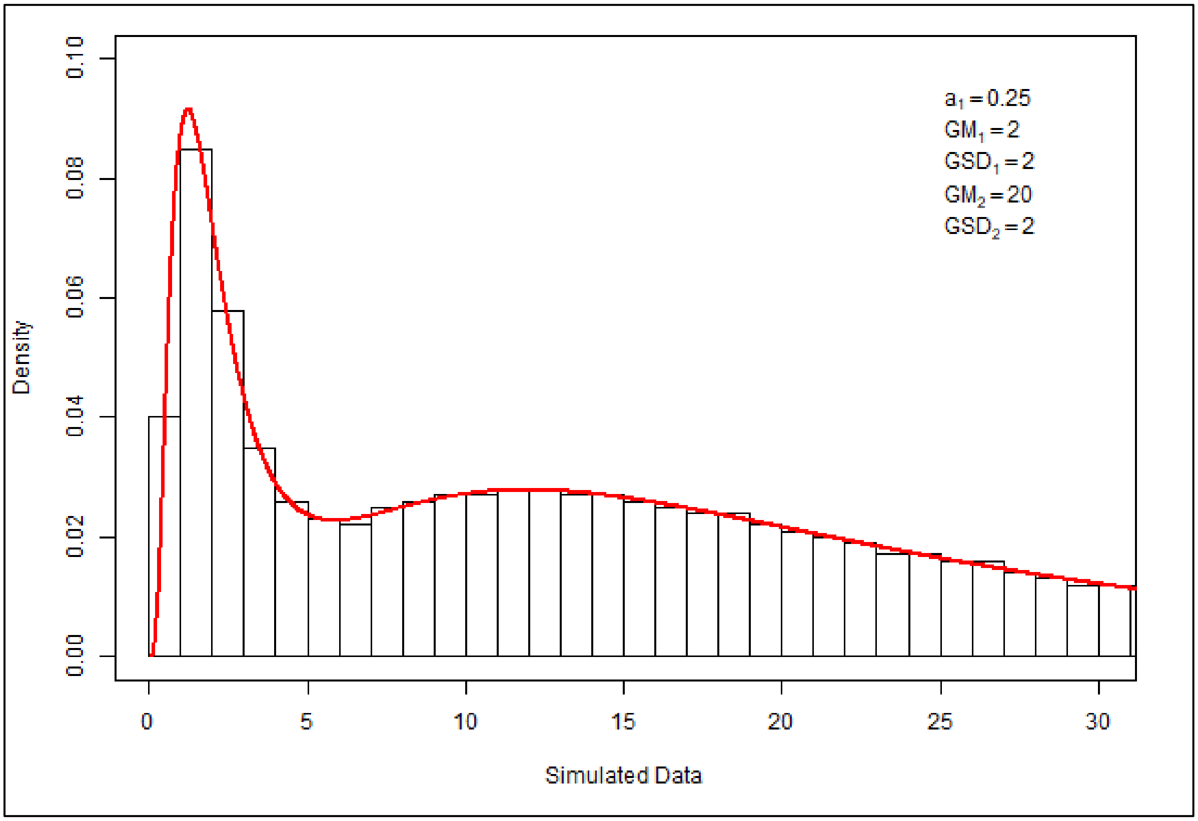

Figure 6-1. Histogram of 1,000 simulated results with true lognormal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

For illustrative purposes, a lognormal distribution was fit to the 1,000 simulated results. Figure 6-2 shows the same histogram from Figure 6-1, but the overlaid curve in Figure 6-2 is the PDF from a lognormal fit. Figure 6-3 is a lognormal probability plot. The data were simulated from a lognormal-lognormal mixture so, unsurprisingly and based on both Figures 6-2 and 6-3, the lognormal distribution does not fit well.

A lognormal-lognormal mixture distribution was fit to the 1,000 simulated results. Figure 6-4 shows the same histogram from Figure 6-1, but the overlaid curve in Figure 6-4 is the PDF from a lognormal-lognormal mixture fit. Figure 6-5 is a probability plot. Unsurprisingly, based on both Figures 6-4 and 6-5, the lognormal-lognormal mixture distribution fits well and the fitted parameters closely match the known parameters for the simulated data.

Table 6-1 summarizes various measures of variability of the true distribution, the simulated data, and the fitted distributions. As one would expect, the lognormal-lognormal mixture column has much better agreement with the true and empirical columns than does the lognormal column.

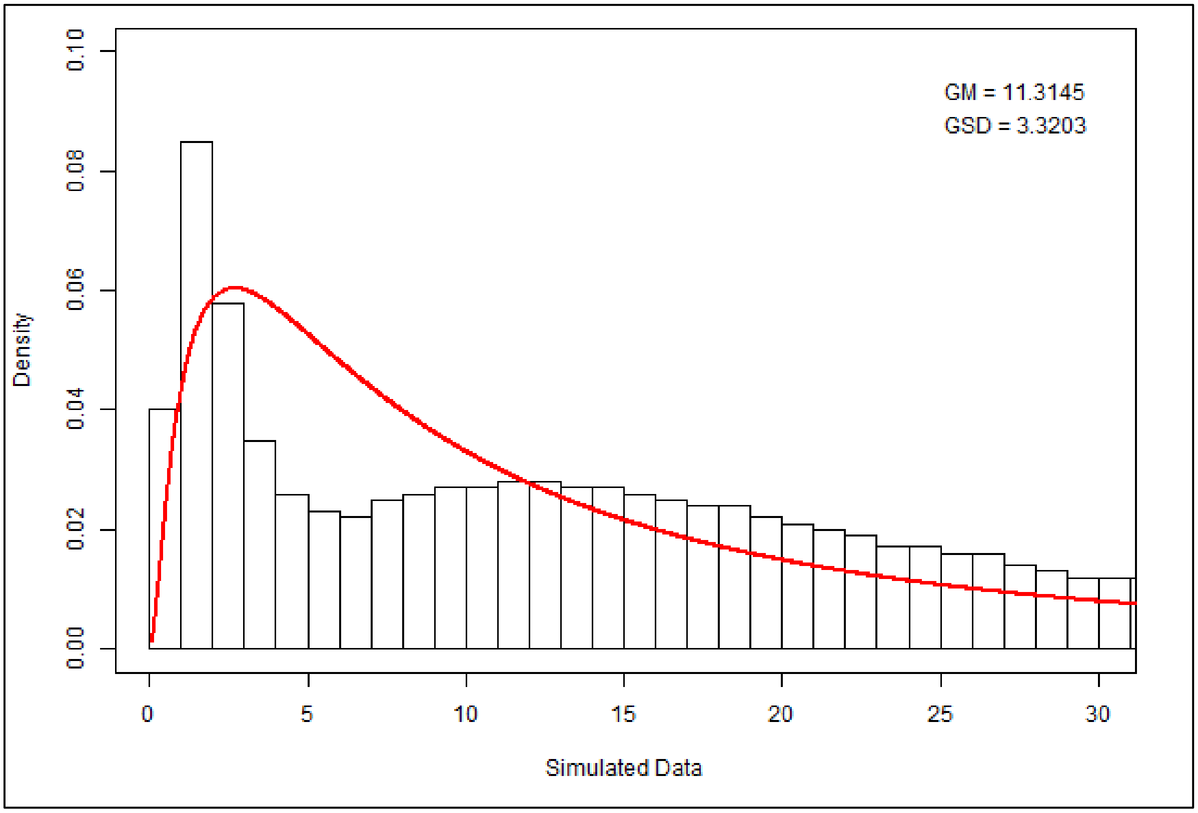

Figure 6-2. Histogram of 1,000 simulated results with fitted lognormal PDF overlaid. Attachment A contains an extended description.

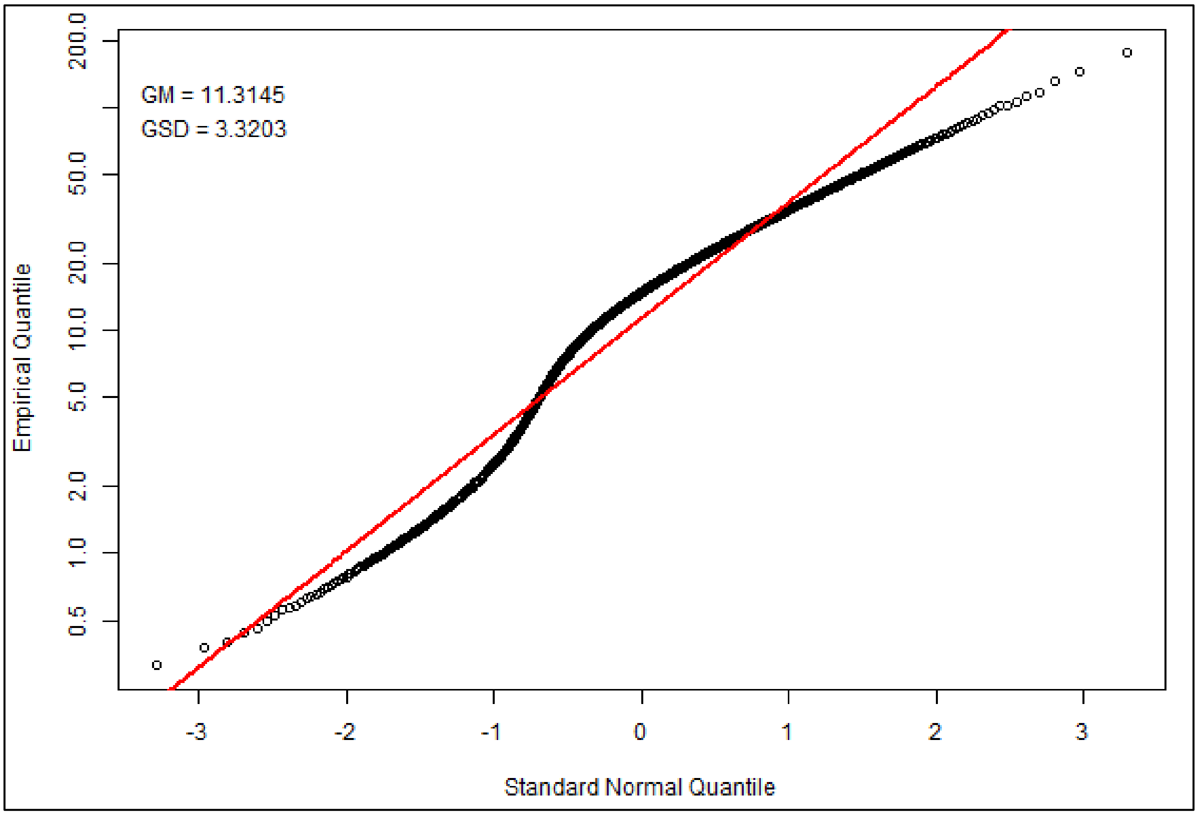

Figure 6-3. Lognormal probability plot of 1,000 simulated results with fitted lognormal. Attachment A contains an extended description.

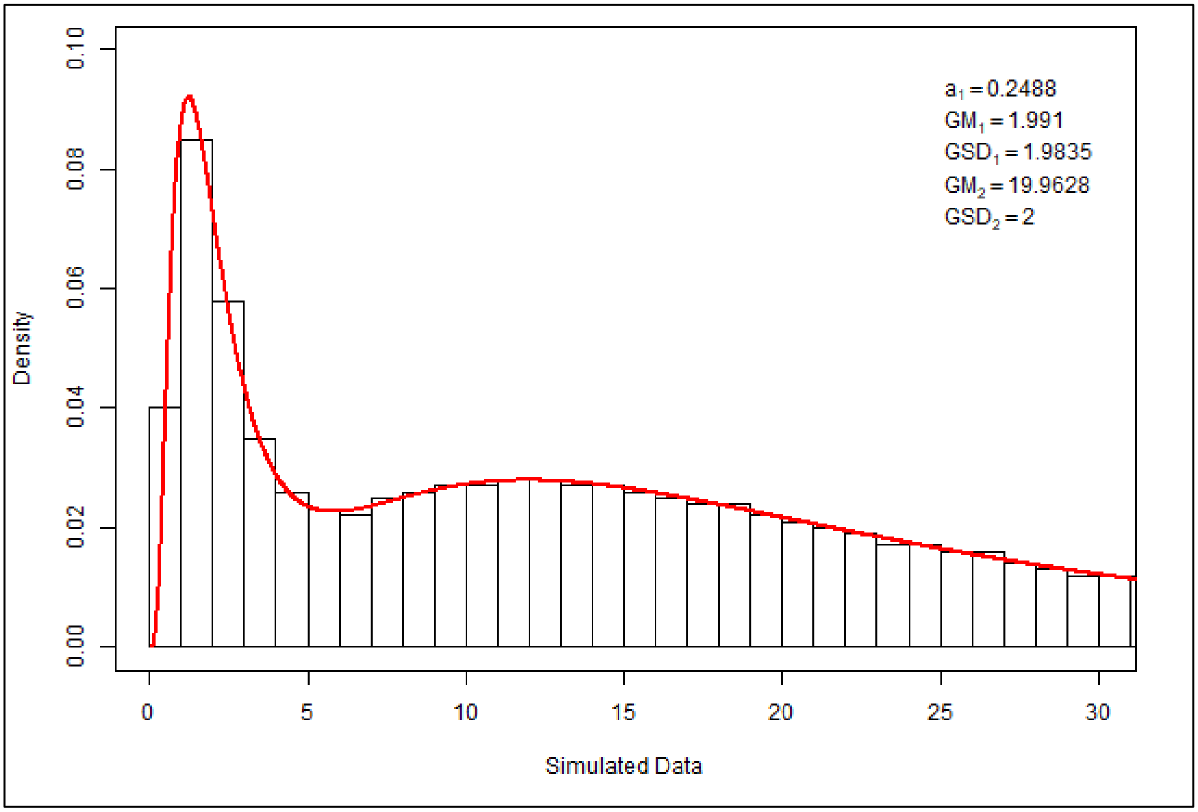

Figure 6-4. Histogram of 1,000 simulated results with fitted lognormal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

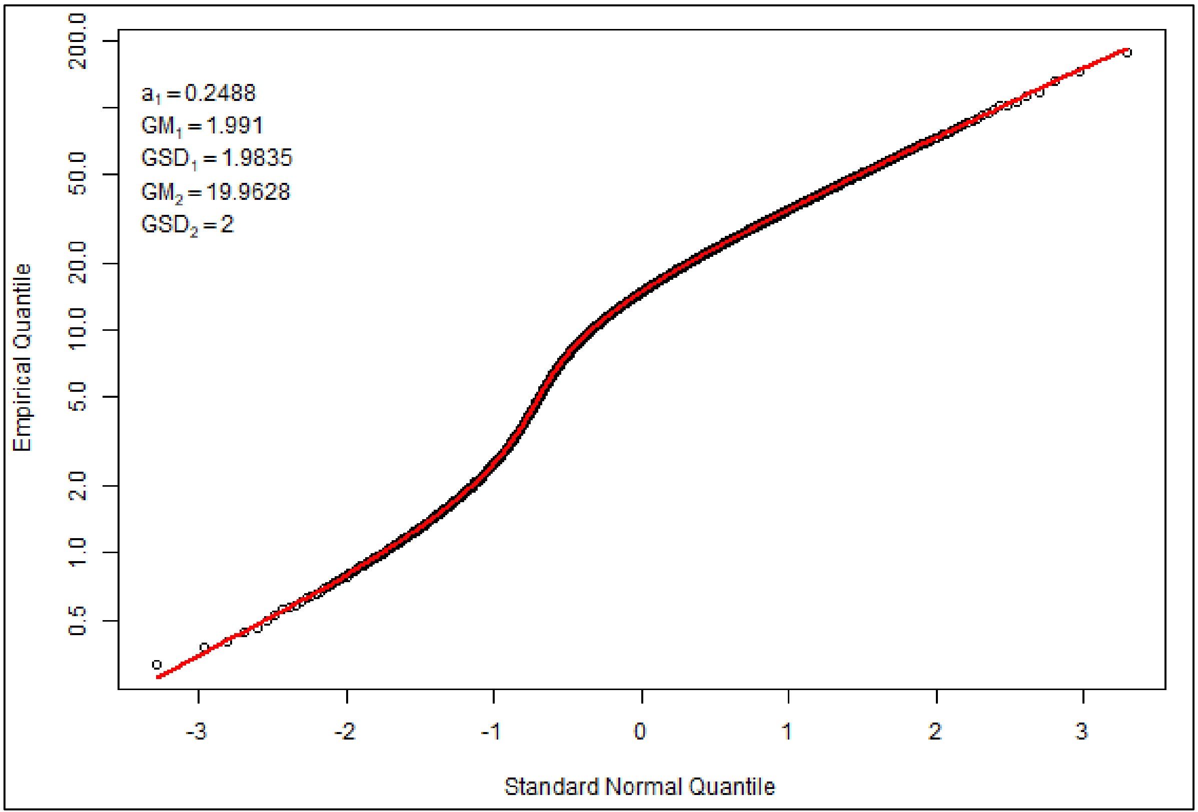

Figure 6-5. Probability plot of 1,000 simulated results with fitted lognormal-lognormal mixture. Attachment A contains an extended description.

Table 6-1. Summary of various measures of variability for lognormal-lognormal mixture distribution.

Measure of variability	True distribution	Simulated data (empirical)	Fitted lognormal	Fitted lognormal-lognormal mixture
Range	(0, ∞)	(0.3146, 175.3741)	(0, ∞)	(0, ∞)
IQR	21.6943	21.6451	20.3831	21.6685
Variance	398.3954	386.5745	1,740.7513	397.2376
Standard deviation	19.9598	19.6615	41.7223	19.9308
$Q_{0.95}/Q_{0.05}$	50.7313	50.4737	51.8228	50.4274
$Q_{0.95}/Q_{0.5}$	3.8107	3.8055	7.1988	3.8060

7.0 HEALTH PHYSICS EXAMPLES NOTE

The examples in Sections 8.0, 9.0, and 10.0 are for illustrative purposes only. They show how a data analyst might attempt to quantify variability, explore stratification, and perhaps fit a mixture distribution. This report does not choose or endorse any analysis. The selection of an analytical approach is the responsibility of a team of statisticians and health physicists, taking into account project consistency and practical applicability. This report does not address the application of these methods, specifically the fitted mixture distributions.

As mentioned in Section 3.2, stratification should be based on what the health physicist knows about the site, processes, and data that would suggest that there are distinct homogeneous subgroups. Because these examples are for illustrative purposes, the stratification analysis in this report only explores possible ways to stratify.

8.0 BIOASSAY DATA EXAMPLE

8.1 DESCRIPTION OF DATA

The bioassay data in this section are Los Alamos National Laboratory (LANL) ^{239}Pu alpha spectrometry results from 2000. This analysis considers 2,152 bioassay results with 1,164 (or 54%) of them being reported as values less than or equal to zero. Figure 8-1 is a histogram of the LANL ^{239}Pu bioassay results.

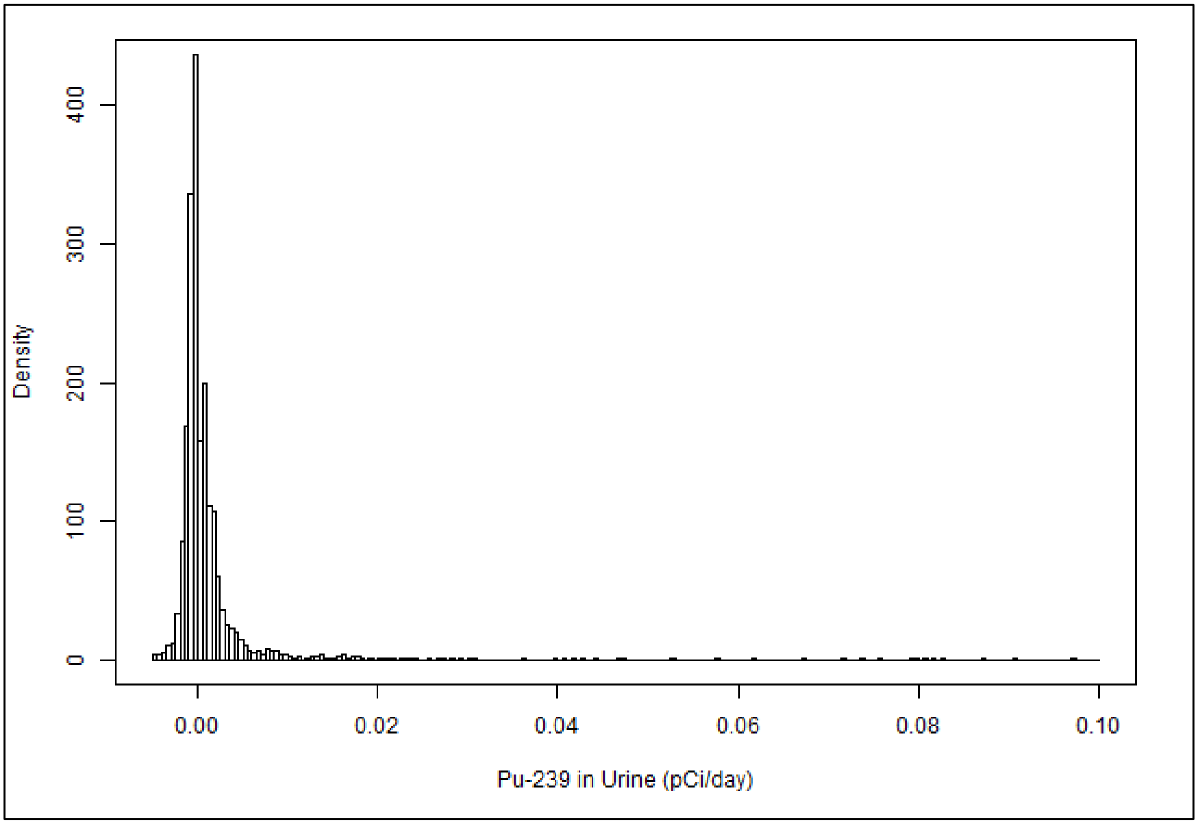

Figure 8-1. Histogram of LANL ^{239}Pu bioassay results.

8.2 LOGNORMAL AND NORMAL FITS

A lognormal distribution was fit to the bioassay results. Figure 8-2 shows the same histogram from Figure 8-1, but the overlaid curve in Figure 8-2 is the PDF from a lognormal fit. Figure 8-3 is a lognormal probability plot. Based on the fits in Figures 8-2 and 8-3, the lognormal distribution does not fit well. The lognormal distribution has no way of handling the bioassay data less than or equal to zero, which is about 54% of the data. Note that the GSD of the lognormal fit is 12.709, which is fairly large.

Because more than half the bioassay data are less than or equal to zero, a normal distribution was also fit to the bioassay results. Figure 8-4 shows the same histogram from Figure 8-1, but the overlaid curve in Figure 8-4 is the PDF from a normal fit. Figure 8-5 is a normal probability plot. The normal distribution does not fit well, mostly because of the influence of the upper tail of the data.

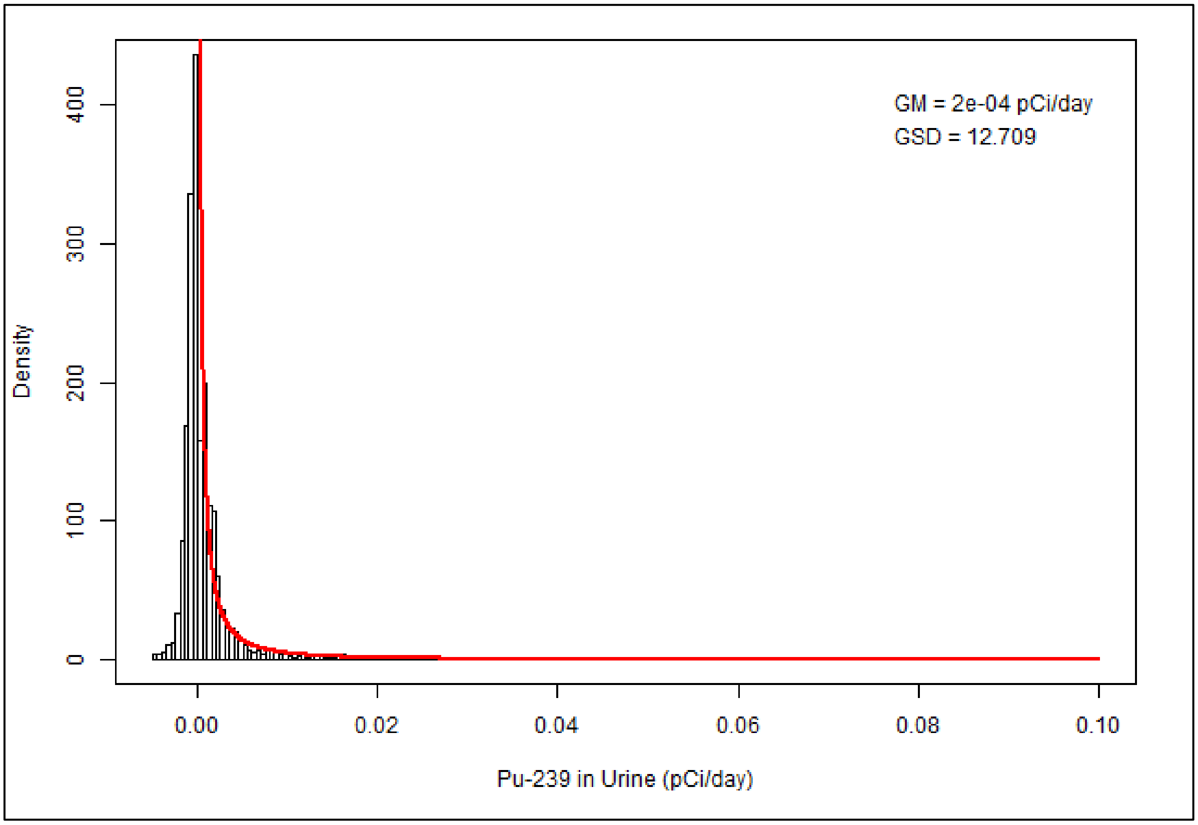

Figure 8-2. Histogram of LANL ^{239}Pu bioassay results with fitted lognormal PDF overlaid.

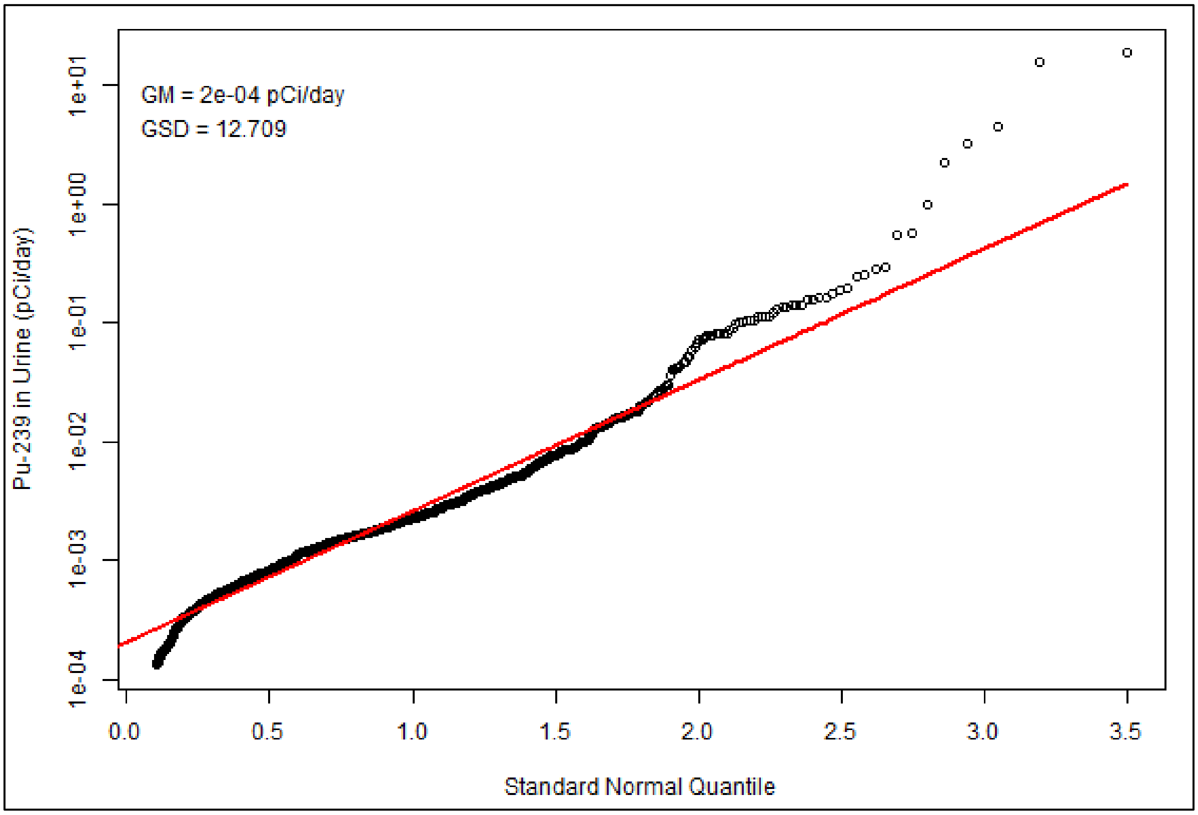

Figure 8-3. Lognormal probability plot of LANL ^{239}Pu bioassay results with fitted lognormal. Attachment A contains an extended description.

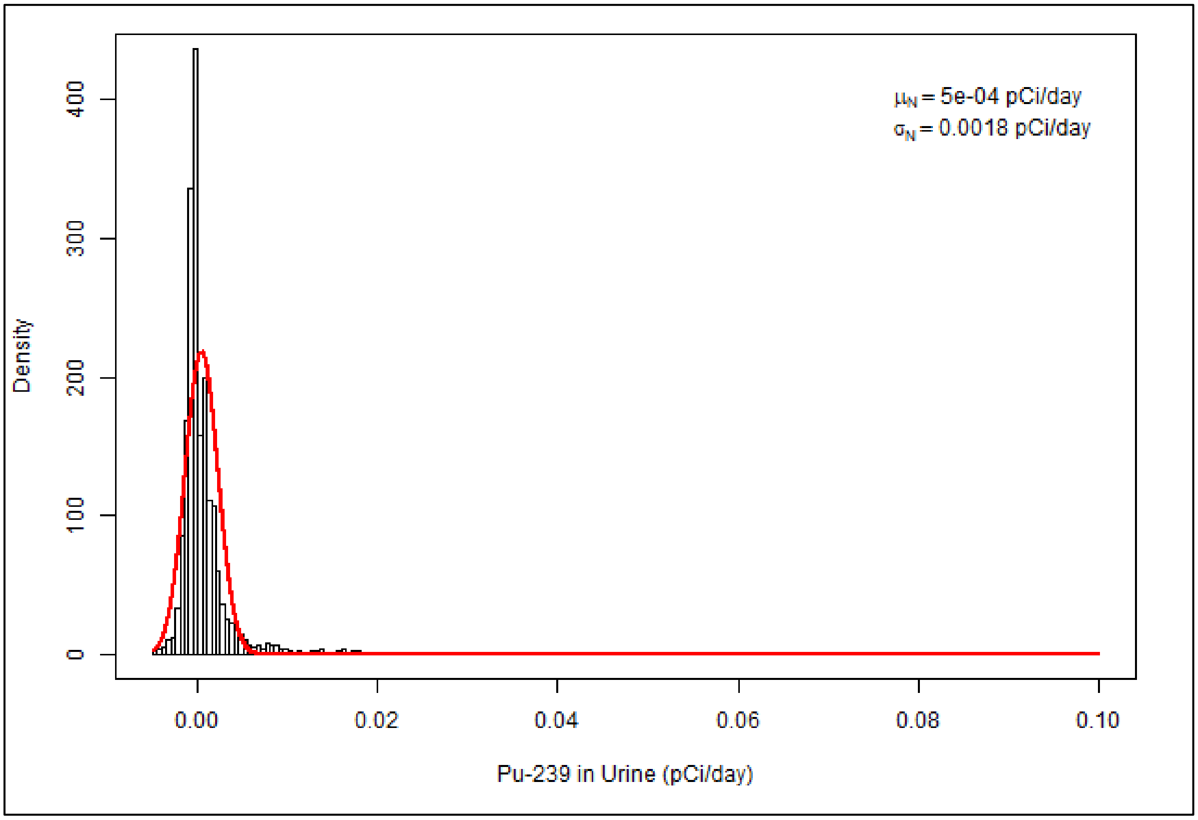

Figure 8-4. Histogram of LANL ^{239}Pu bioassay results with fitted normal PDF overlaid.

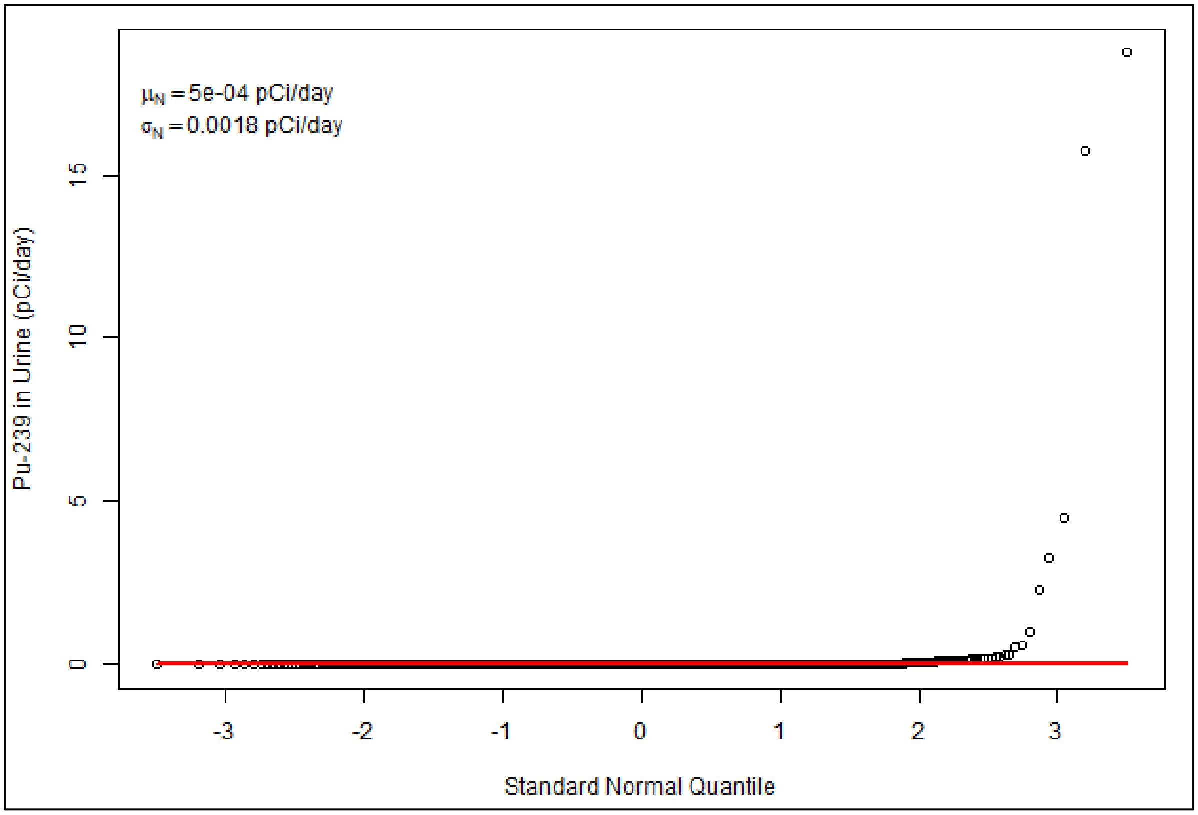

Figure 8-5. Normal probability plot of LANL ^{239}Pu bioassay results with fitted normal. Attachment A contains an extended description.

8.3 STRATIFICATION

In the dataset from which the LANL ^{239}Pu data were extracted, the only possible stratification variable is the schedule type: routine, special, or prompt action. Figure 8-6 is a four-panel plot that includes the histogram from Figure 8-1 as the upper-left panel. The other three panels are histograms (on the same *x*-scale) of the three schedule types. Figure 8-7 contains side-by-side boxplots for each of the three schedule types.

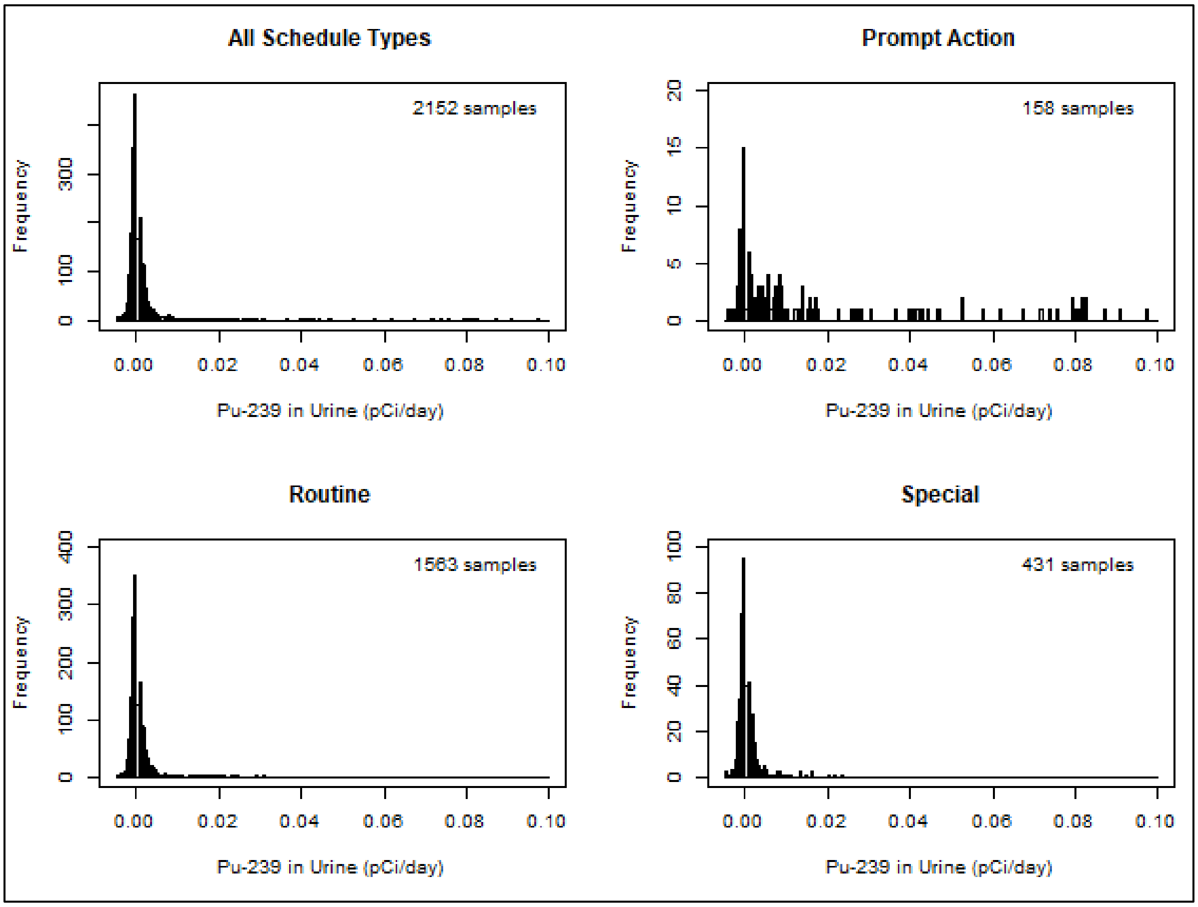

Figure 8-6. Four-panel plot of histograms of all schedule types, prompt action, routine, and special samples. Attachment A contains an extended description.

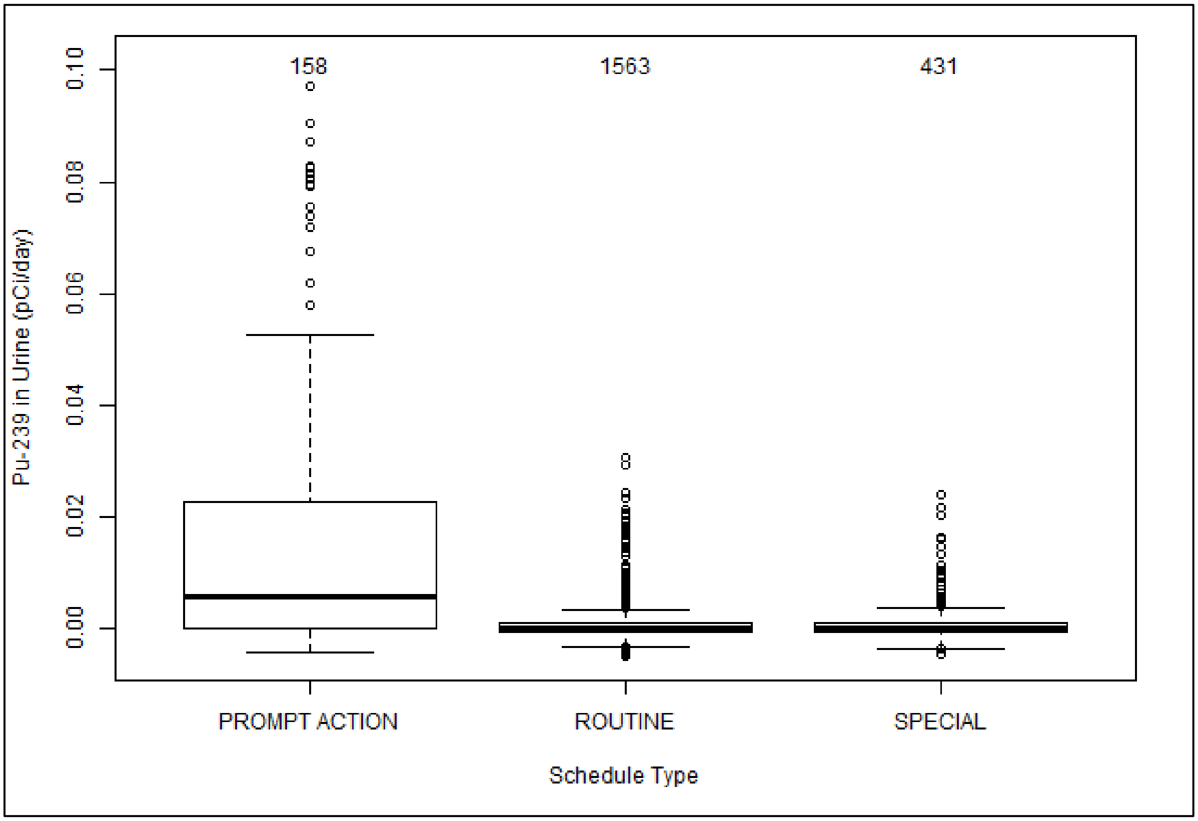

Figure 8-7. Side-by-side boxplots for each schedule type: prompt action, routine, and special samples. Attachment A contains an extended description.

From Figures 8-6 and 8-7, the empirical distribution of routines looks similar to the empirical distribution of specials, and the empirical distribution of prompt action samples looks different from the other two sample types. Therefore, the subsequent analysis examines prompt action samples separately from routines and specials.

8.3.1 Routine and Special Samples

A lognormal distribution was fit to the routine and special results. Figure 8-8 is a histogram with the overlaid PDF from a lognormal fit. Figure 8-9 is a lognormal probability plot. The lognormal fit is adequate, but it still has no way of handling the data less than or equal to zero. Note that the GSD of the lognormal fit is 6.3608, which is about half of the GSD from Figures 8-2 and 8-3.

A normal distribution was also fit to the bioassay results. Figure 8-10 is a histogram with the overlaid PDF from a normal fit. Figure 8-11 is a normal probability plot. The normal fit does a decent job except right around zero and the extreme upper tail.

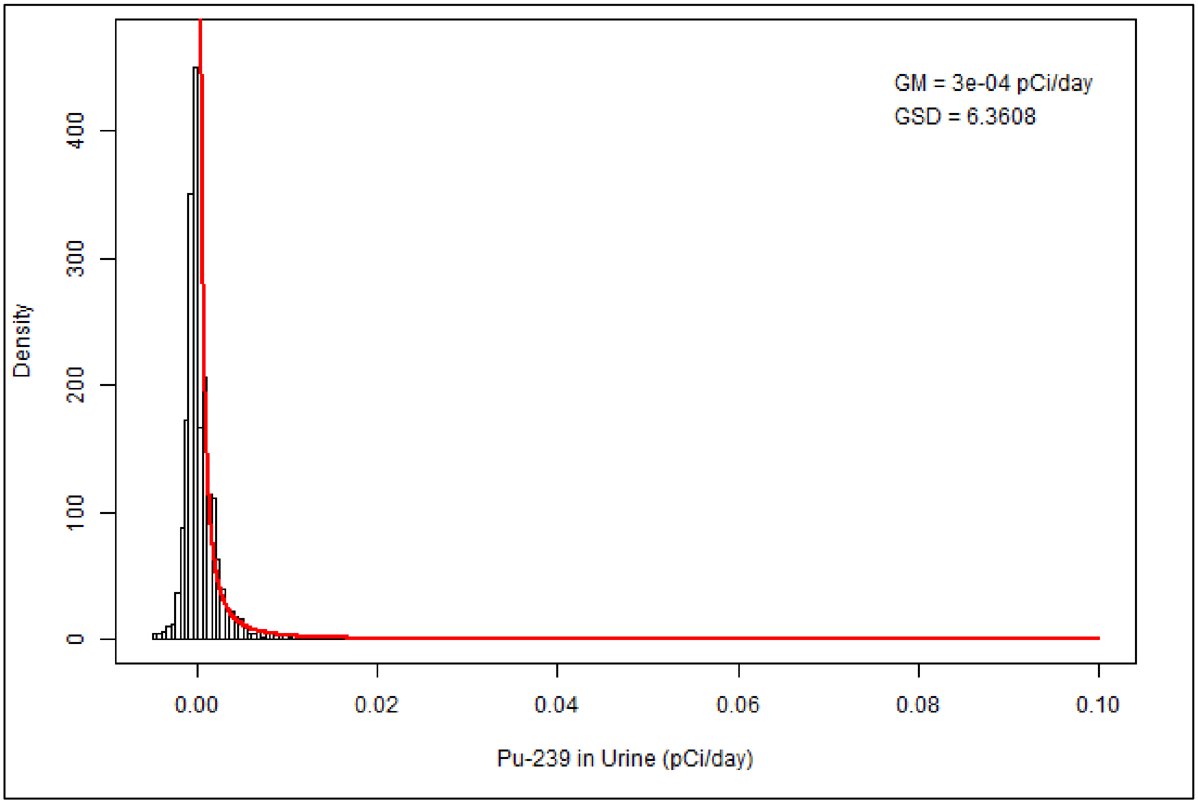

Figure 8-8. Histogram of routine and special LANL ^{239}Pu bioassay results with fitted lognormal PDF overlaid. Attachment A contains an extended description.

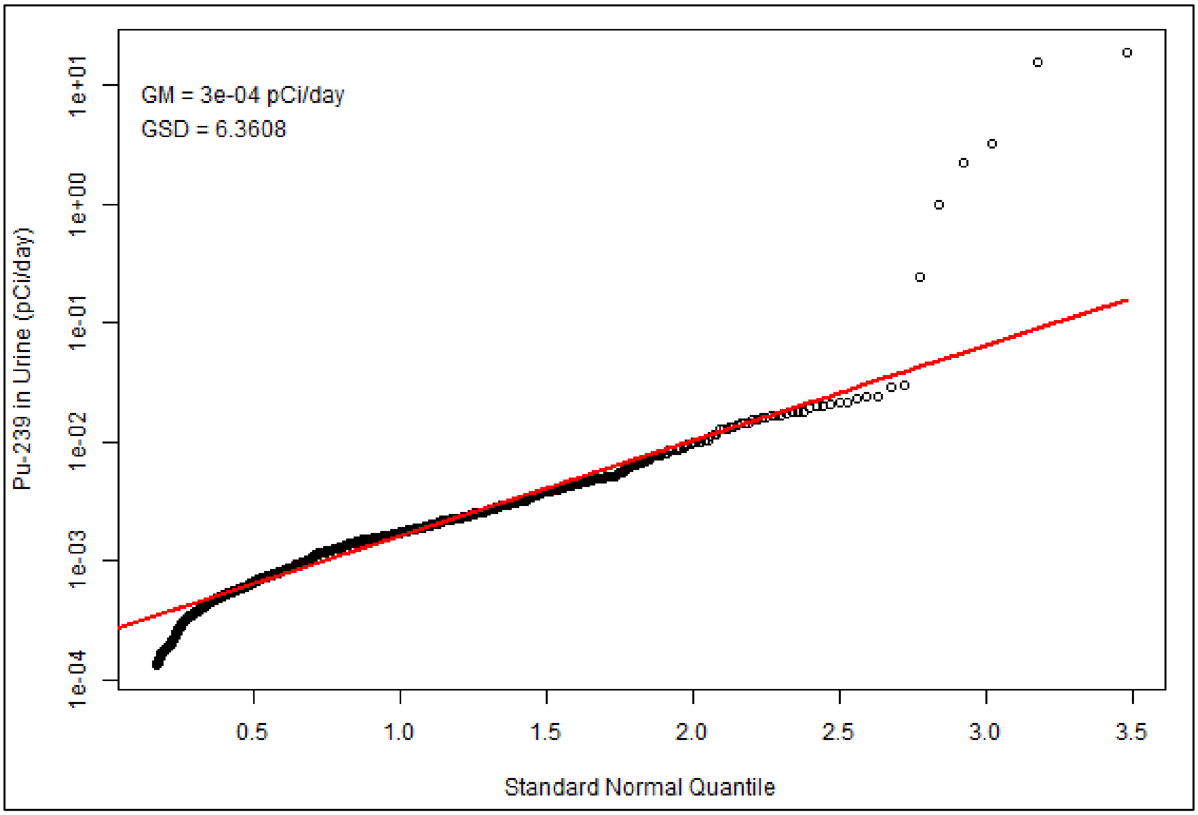

Figure 8-9. Lognormal probability plot of routine and special LANL ^{239}Pu bioassay results with fitted lognormal. Attachment A contains an extended description.

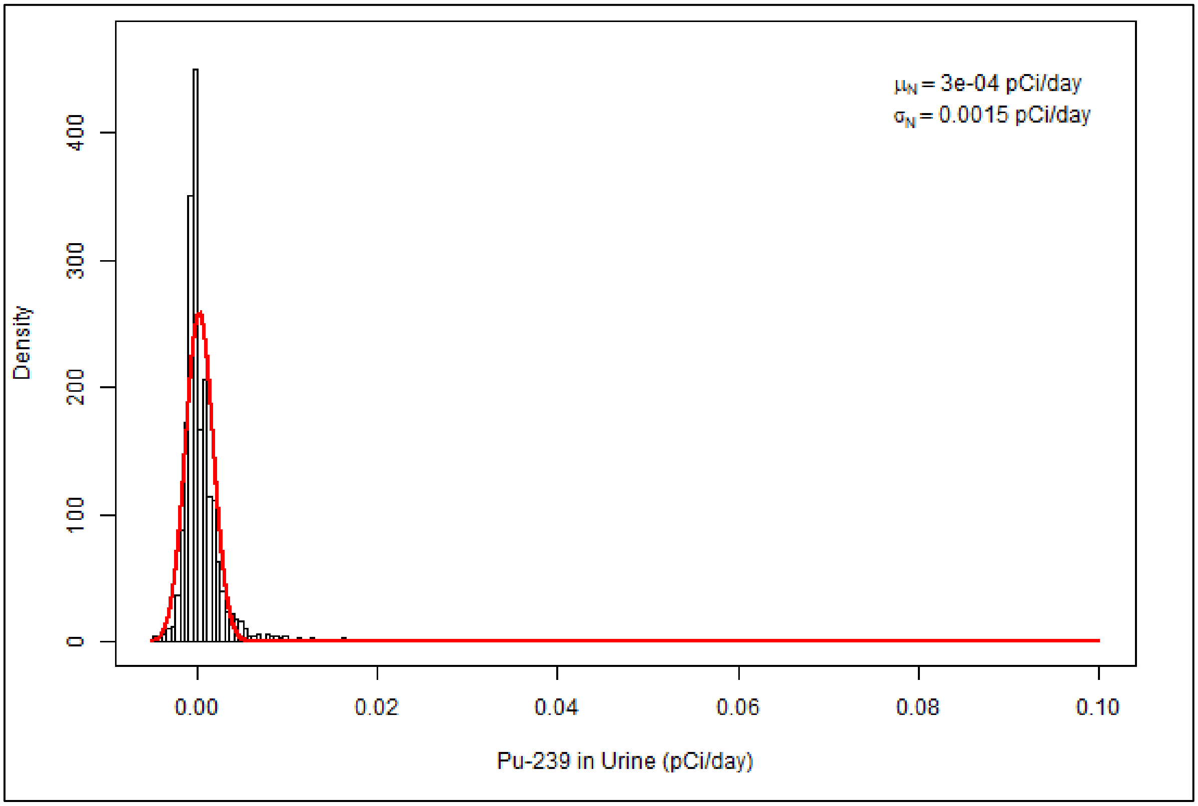

Figure 8-10. Histogram of routine and special LANL ^{239}Pu bioassay results with fitted normal PDF overlaid. Attachment A contains an extended description.

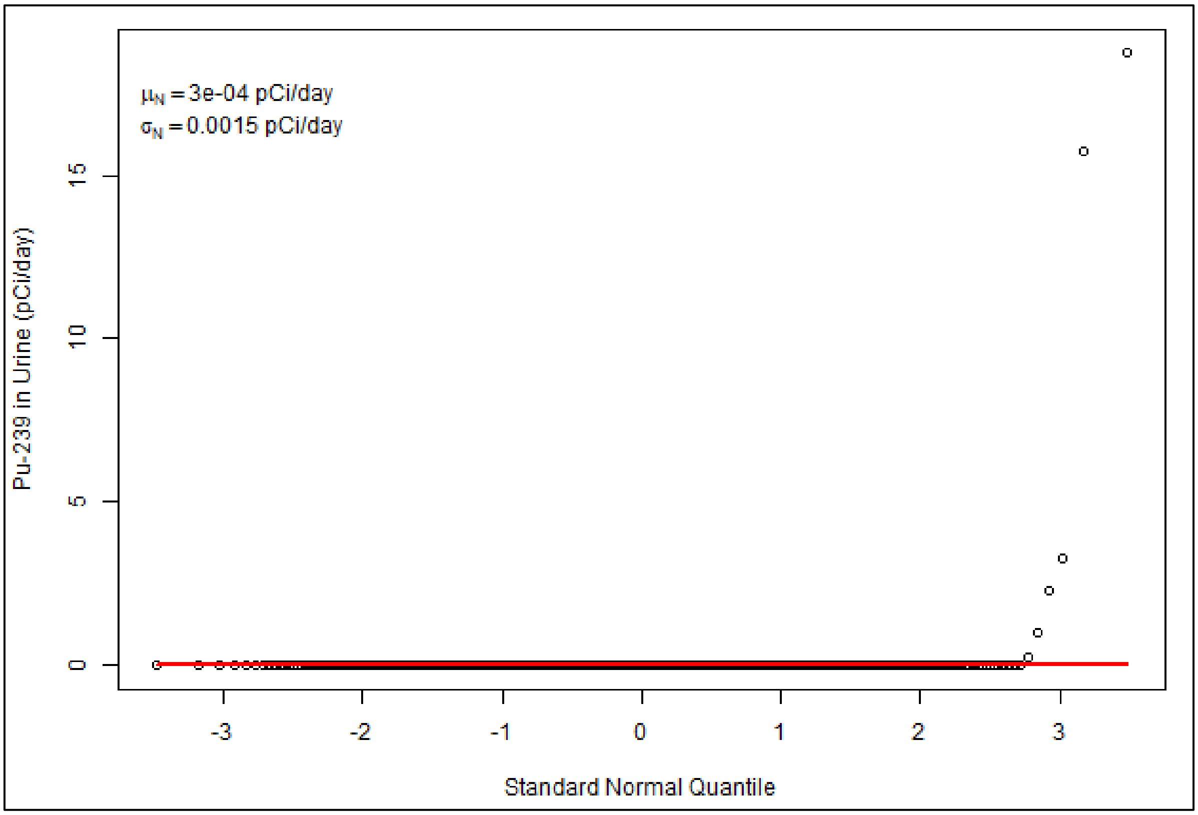

Figure 8-11. Normal probability plot of routine and special LANL ^{239}Pu bioassay results with fitted normal. Attachment A contains an extended description.

8.3.2 Prompt Action Samples

A lognormal distribution was fit to the prompt action results. Figure 8-12 is a histogram with the overlaid PDF from a lognormal fit. Figure 8-13 is a lognormal probability plot. The lognormal fit is questionable, and it has no way of handling the data less than or equal to zero.

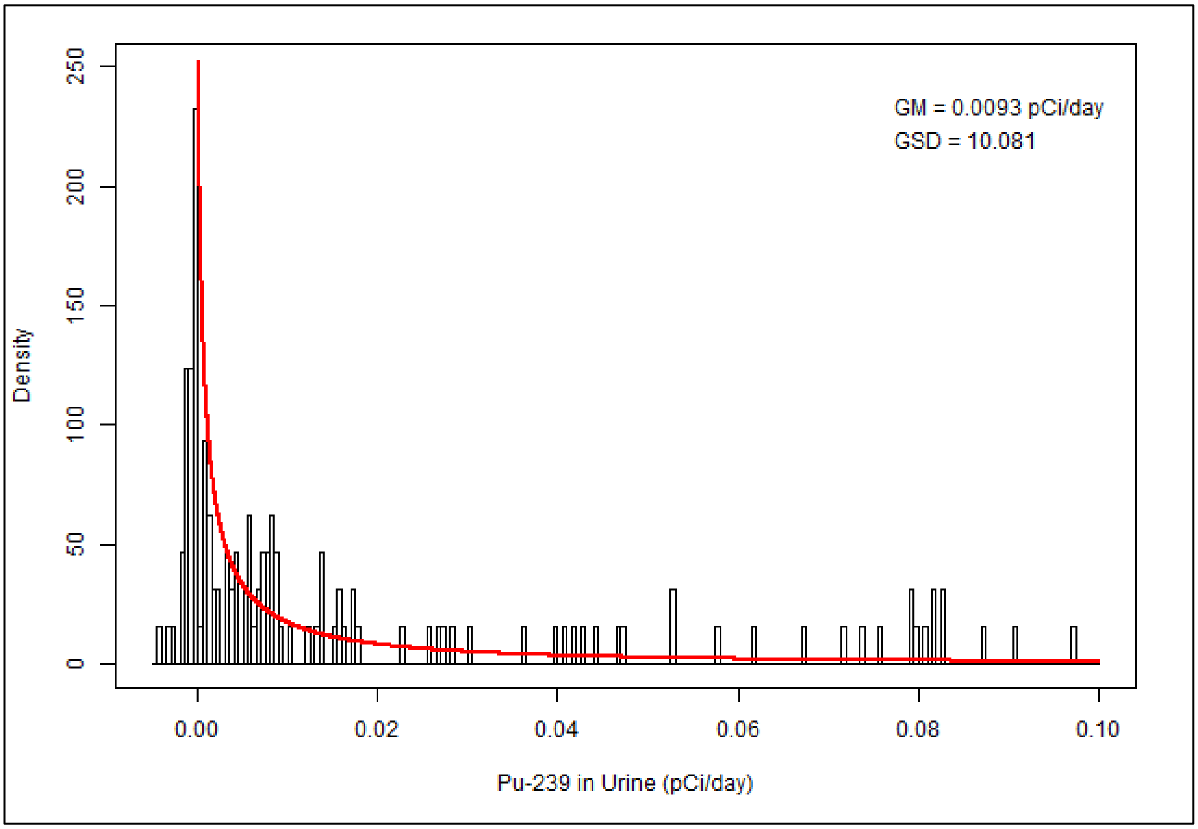

Figure 8-12. Histogram of prompt action LANL ^{239}Pu bioassay results with fitted lognormal PDF overlaid. Attachment A contains an extended description.

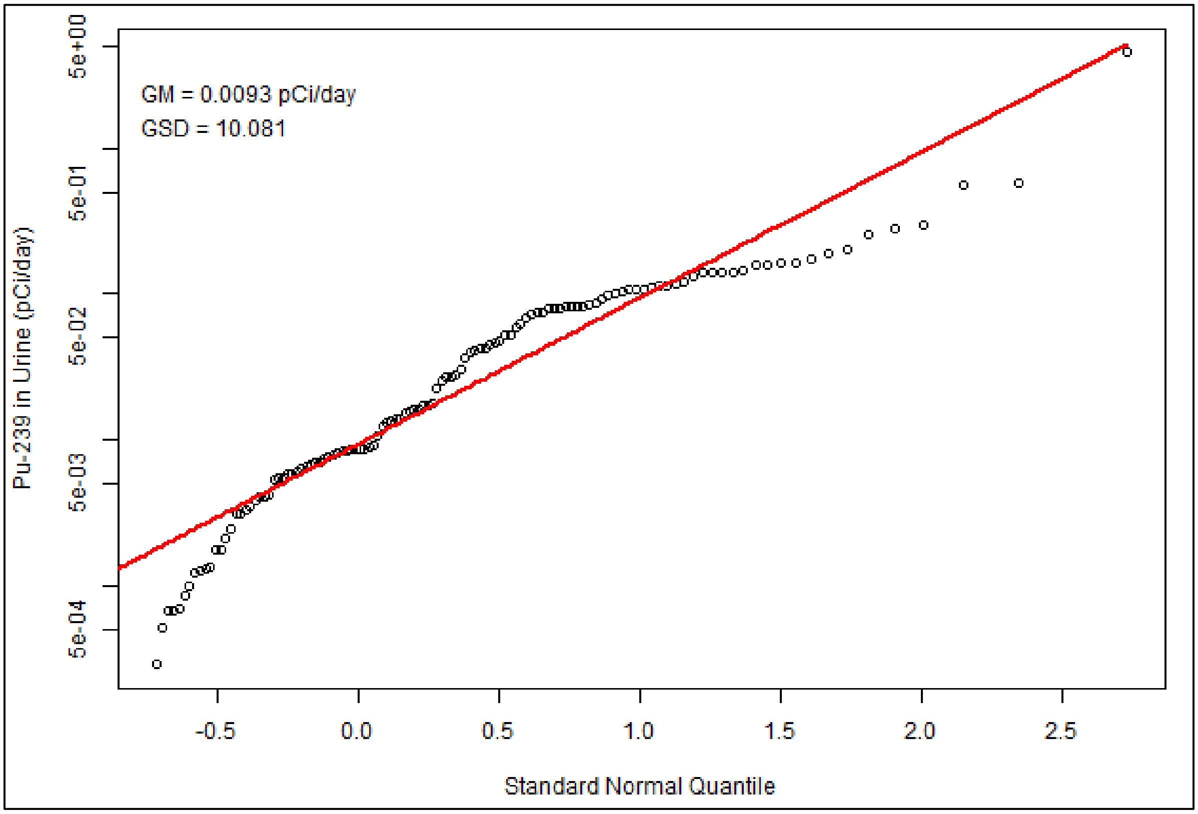

Figure 8-13. Lognormal probability plot of prompt action LANL ^{239}Pu bioassay results with fitted lognormal. Attachment A contains an extended description.

If these stratified data were actually being analyzed for use, these stratified groups could be fit with other distributions (perhaps fitting mixture distributions to the stratified data) to improve the fits. The schedule type was the only available stratification variable, but if the data were being analyzed for use, ancillary stratification information could be considered. Because these examples are only for illustrative purposes, the analysis of the stratified groups ends here.

8.4 NORMAL-LOGNORMAL MIXTURE FIT

A normal-lognormal mixture distribution was fit to the unstratified LANL ^{239}Pu bioassay results. Figure 8-14 shows the same histogram from Figure 8-1, but the overlaid curve in Figure 8-14 is the PDF from a normal-lognormal mixture fit. Figure 8-15 is a probability plot. Based on both Figures 8-14 and 8-15, the normal-lognormal mixture distribution fits well. The fitted parameters indicate that about 83% of the distribution is normal and that the normal tends to represent the lower portion of the data with a mean of essentially zero. That would suggest that the majority of the results are normal noise (centered at zero) with some lognormal signal. Note that the GSD of the lognormal portion of the mixture fit is 5.7699, which is an improvement over the GSD of 12.709 from Figures 8-2 and 8-3.

Table 8-1 summarizes various measures of variability of the data and the distributions fit to the unstratified data. The normal-lognormal mixture column has better agreement with the empirical column (specifically the robust summary statistics: IQR and $Q_{0.95}/Q_{0.05}$) than do the lognormal and normal columns.

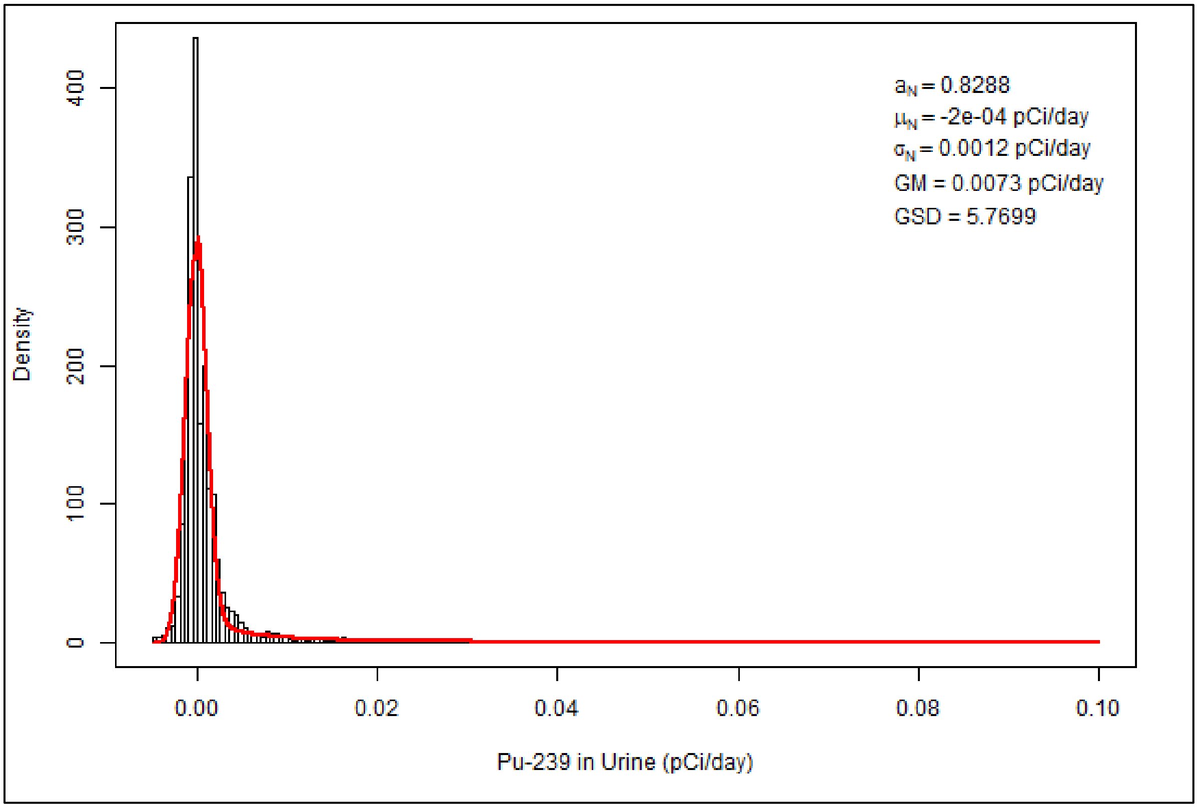

Figure 8-14. Histogram of LANL ^{239}Pu bioassay results with fitted normal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

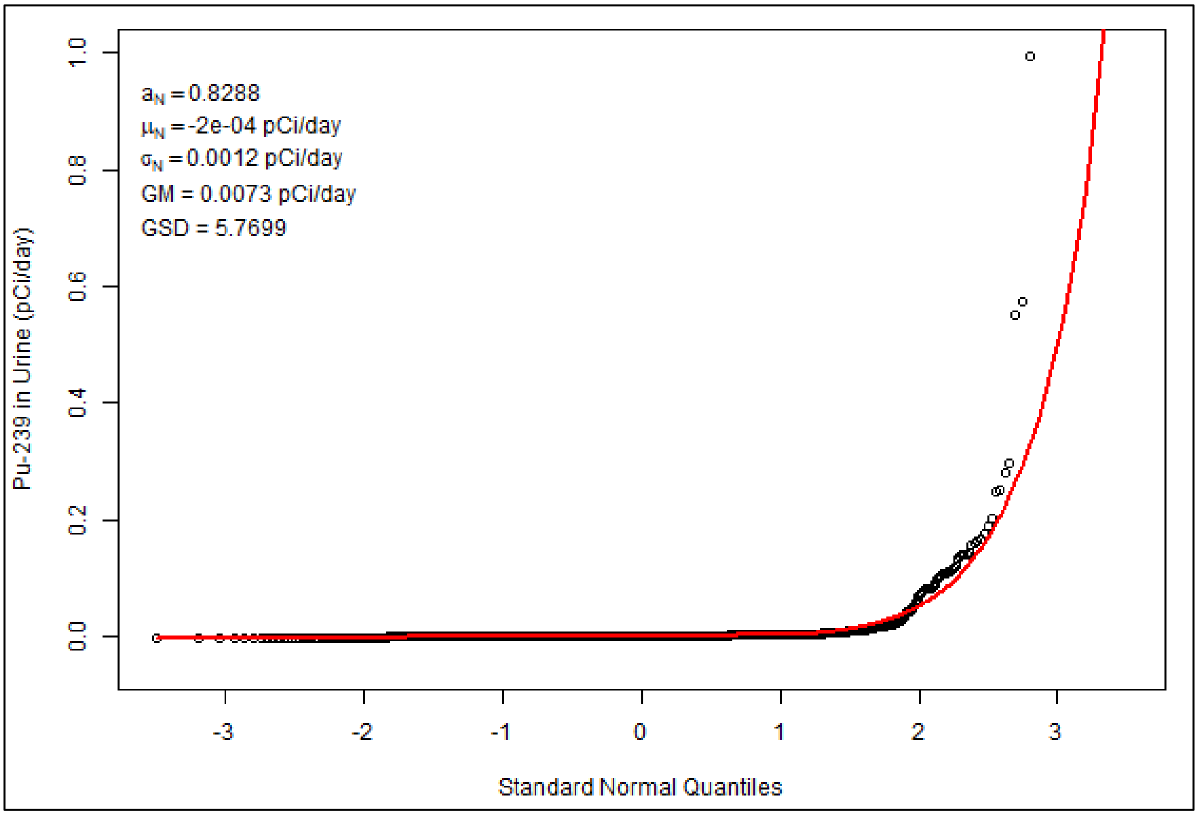

Figure 8-15. Probability plot of LANL ^{239}Pu bioassay results with fitted normal-lognormal mixture. Attachment A contains an extended description.

Table 8-1. Summary of various measures of variability for LANL ^{239}Pu bioassay results.

Measure of variability	Empirical	Fitted lognormal	Fitted normal	Fitted normal-lognormal mixture
Range	(−0.0049, 18.74)	(0, ∞)	(-∞, ∞)	(-∞, ∞)
IQR	0.0020	0.0011	0.0025	0.0020
Variance	0.2955	0.0172	3.3433E−6	0.0042
Standard deviation	0.5436	0.1311	0.0018	0.0648
$Q_{0.95}/Q_{0.05}$	−7.3631	4,287.4766	−1.3953	−9.4182
$Q_{0.95}/Q_{0.5}$	Undefined[a]	65.4788	7.0592	128.4344

a. The empirical 50th percentile is zero, so this ratio divides by zero.

9.0 DOSIMETER DATA EXAMPLE

9.1 DESCRIPTION OF DATA

The dosimeter data in this section are Y-12 Plant deep dose results from 1953. In this dataset, despite the exchange cycle, results are recorded quarterly. This analysis considers 659 quarterly doses, with 253 (or 38%) of them being reported as zero, which means they were censored at 50 mrem.

Annual dose distributions are typically of interest for external dosimetry rather than the quarterly recorded doses in the Y-12 dataset. For this analysis, quarterly doses that were recorded as zero were imputed (single imputation) before annual doses were calculated. Details of the imputation can be found in ORAUT-RPRT-0071, *External Dose Coworker Methodology* [ORAUT 2015, pp. 7–10] and the support files [ORAUT 2024]. The remainder of this section focuses on analyzing Y-12 annual deep doses from 1953. Figure 9-1 is a histogram of the annual deep doses.

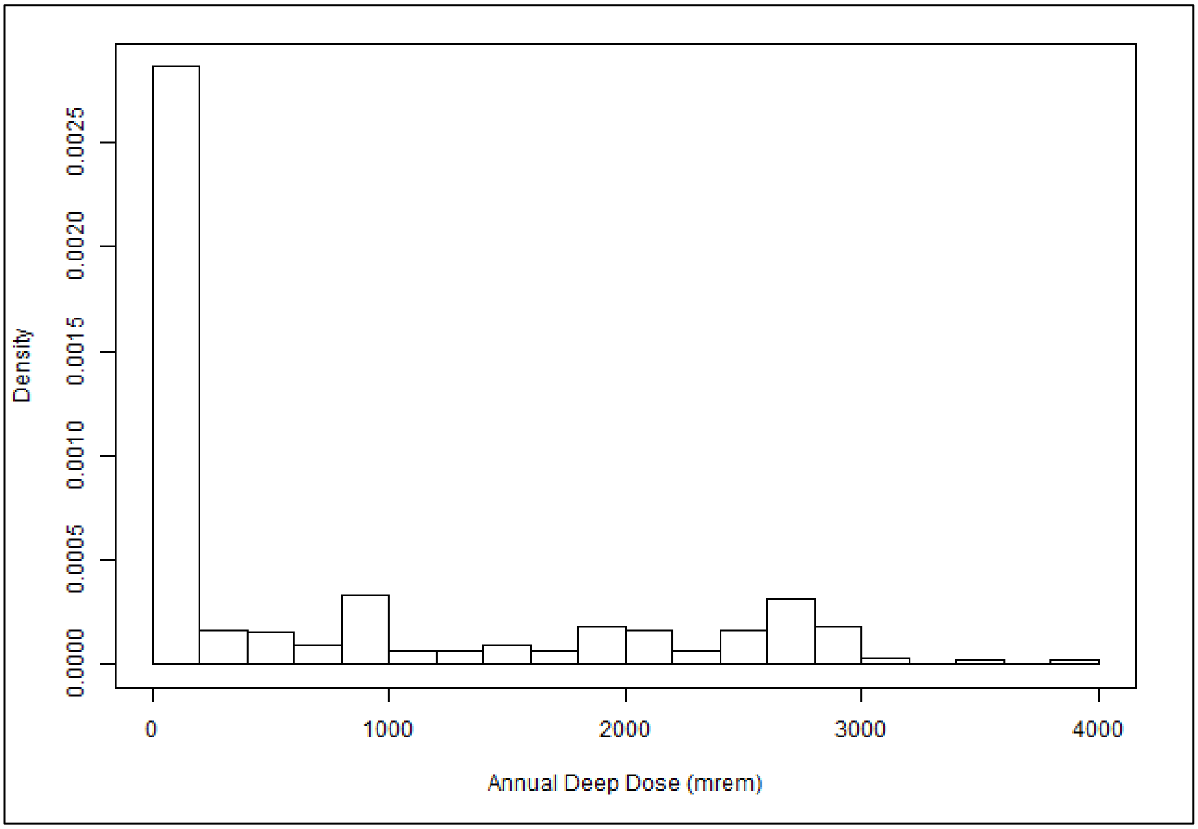

Figure 9-1. Histogram of Y-12 annual deep doses. Attachment A contains an extended description.

9.2 LOGNORMAL AND NORMAL FITS

A lognormal distribution was fit to the annual deep doses. Figure 9-2 shows the same histogram from Figure 9-1, but the overlaid curve in Figure 9-2 is the PDF from a lognormal fit. Figure 9-3 is a lognormal probability plot. The lognormal distribution does not fit well. The lognormal distribution has no way of handling what appears to be the effect of an administrative limit. Note that the GSD of the lognormal fit is 12.9777, which is fairly large.

A normal distribution was also fit to the annual deep doses. Figure 9-4 shows the same histogram from Figure 9-1, but the overlaid curve in Figure 9-4 is the PDF from a normal fit. Figure 9-5 is a normal probability plot. The normal distribution does not fit well.

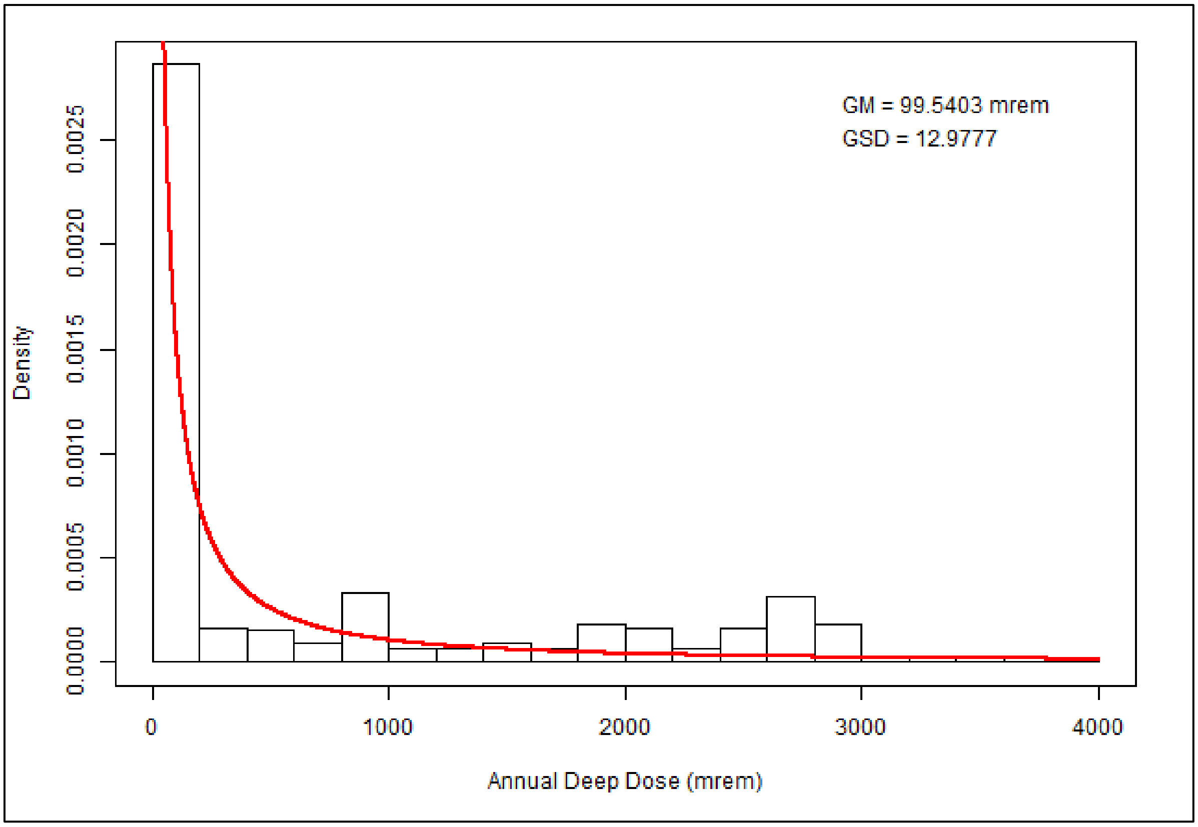

Figure 9-2. Histogram of Y-12 annual deep doses with fitted lognormal PDF overlaid.

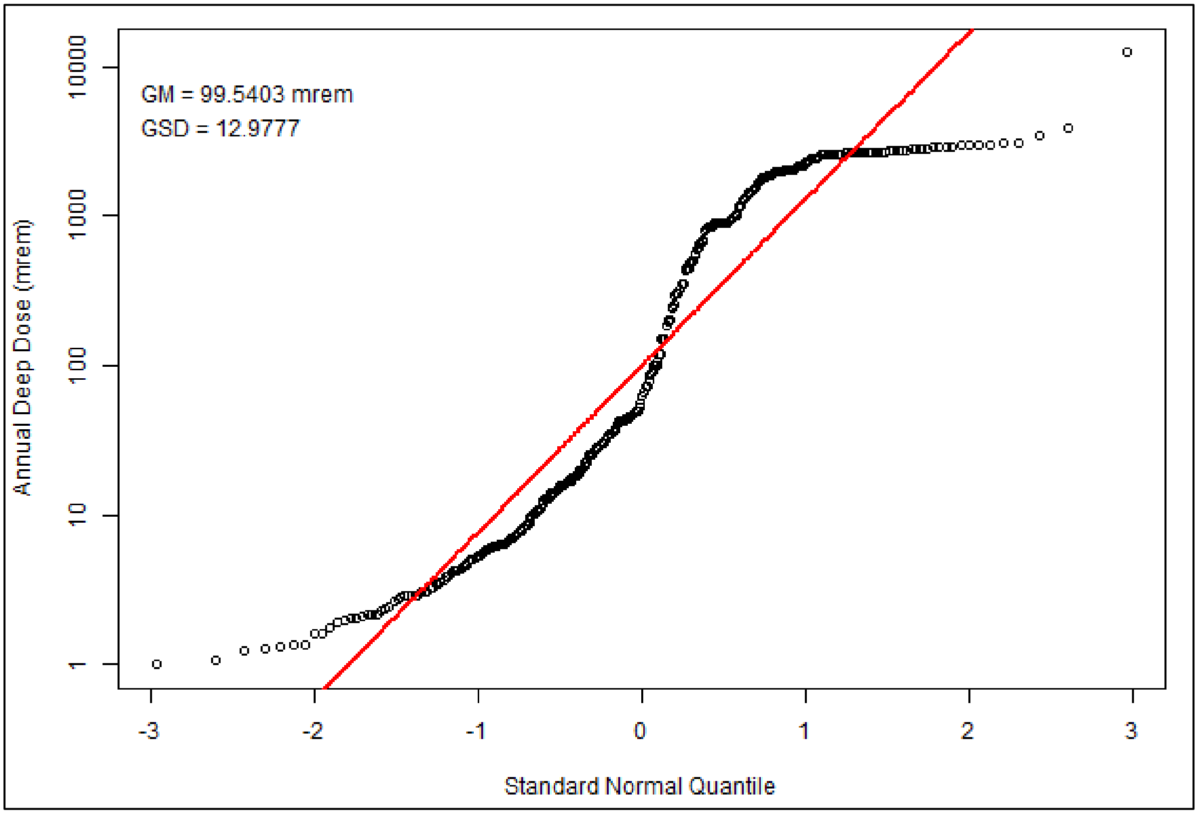

Figure 9-3. Lognormal probability plot of Y-12 annual deep doses with fitted lognormal. Attachment A contains an extended description.

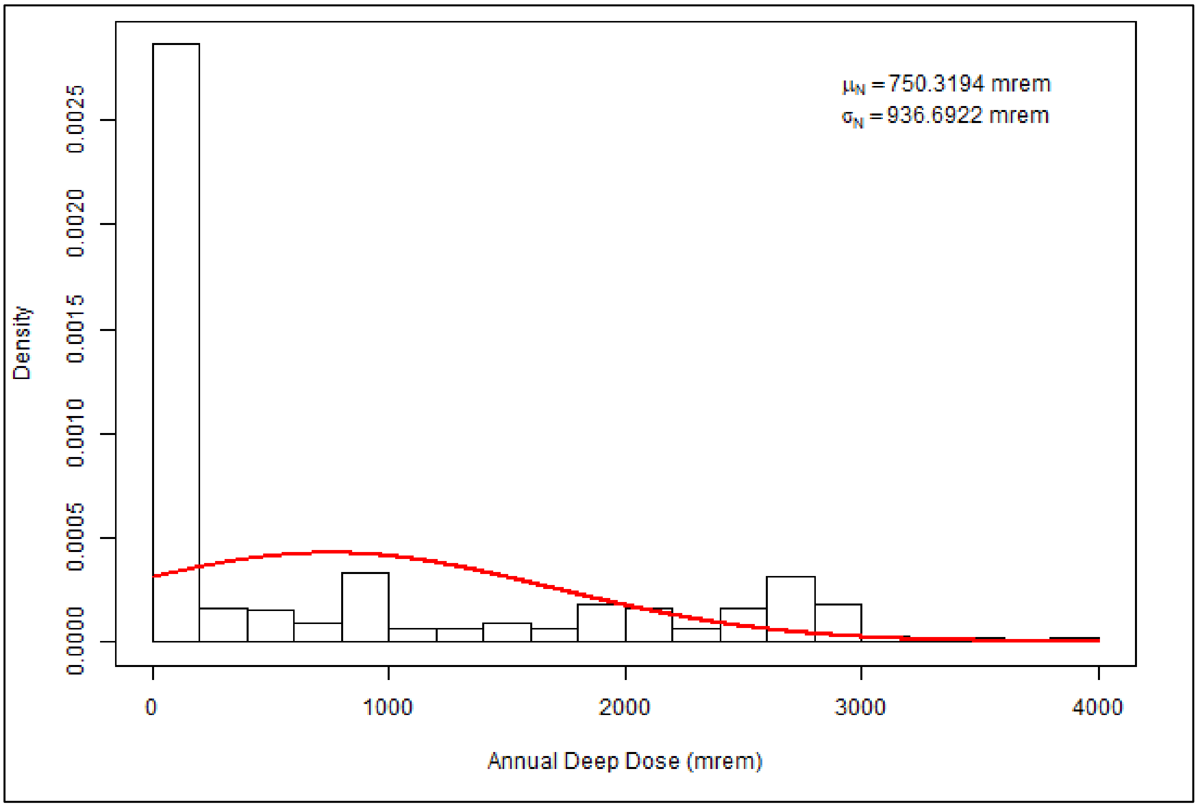

Figure 9-4. Histogram of Y-12 annual deep doses with fitted normal PDF overlaid. Attachment A contains an extended description.

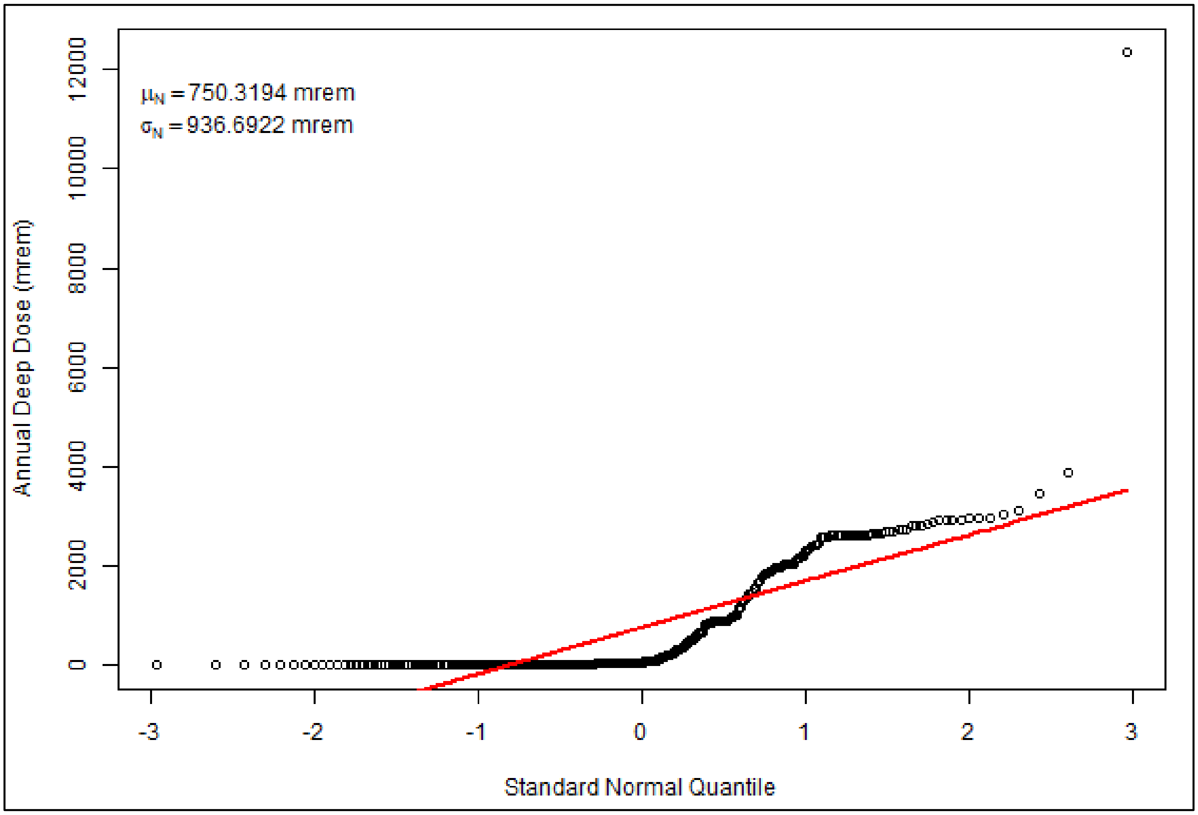

Figure 9-5. Normal probability plot of Y-12 annual deep doses with fitted normal. Attachment A contains an extended description.

9.3 STRATIFICATION

In the dataset from which the Y-12 quarterly deep doses were extracted, the only possible stratification variable is the department code. All results from 1953 are recorded with department code 2000, so no stratification is possible using information in the dataset. As mentioned in Section 3.2, information in an ancillary source can also be considered for stratification, but those sources were not considered for this report.

9.4 NORMAL-LOGNORMAL MIXTURE FIT

A normal-lognormal mixture distribution was fit to the Y-12 annual deep doses. Figure 9-6 shows the same histogram from Figure 9-1, but the overlaid curve in Figure 9-6 is the PDF from a normal-lognormal mixture fit. Figure 9-7 is a probability plot. Based on both Figures 9-6 and 9-7, the normal-lognormal mixture distribution fits well. The fitted parameters indicate that about 38% of the distribution is normal and that the normal tends to represent the upper portion of the data. That would suggest that the majority of the results are lognormally distributed doses with the larger annual doses being affected by the administrative limit, which the normal portion captures fairly well. Note that the GSD of the lognormal portion of the mixture fit is 5.2201, which is an improvement over the GSD of 12.9777 from Figures 9-2 and 9-3.

Table 9-1 summarizes various measures of variability of the data and the distributions fit to the annual dose data. The normal-lognormal mixture column has better agreement with the empirical column than do the lognormal and normal columns.

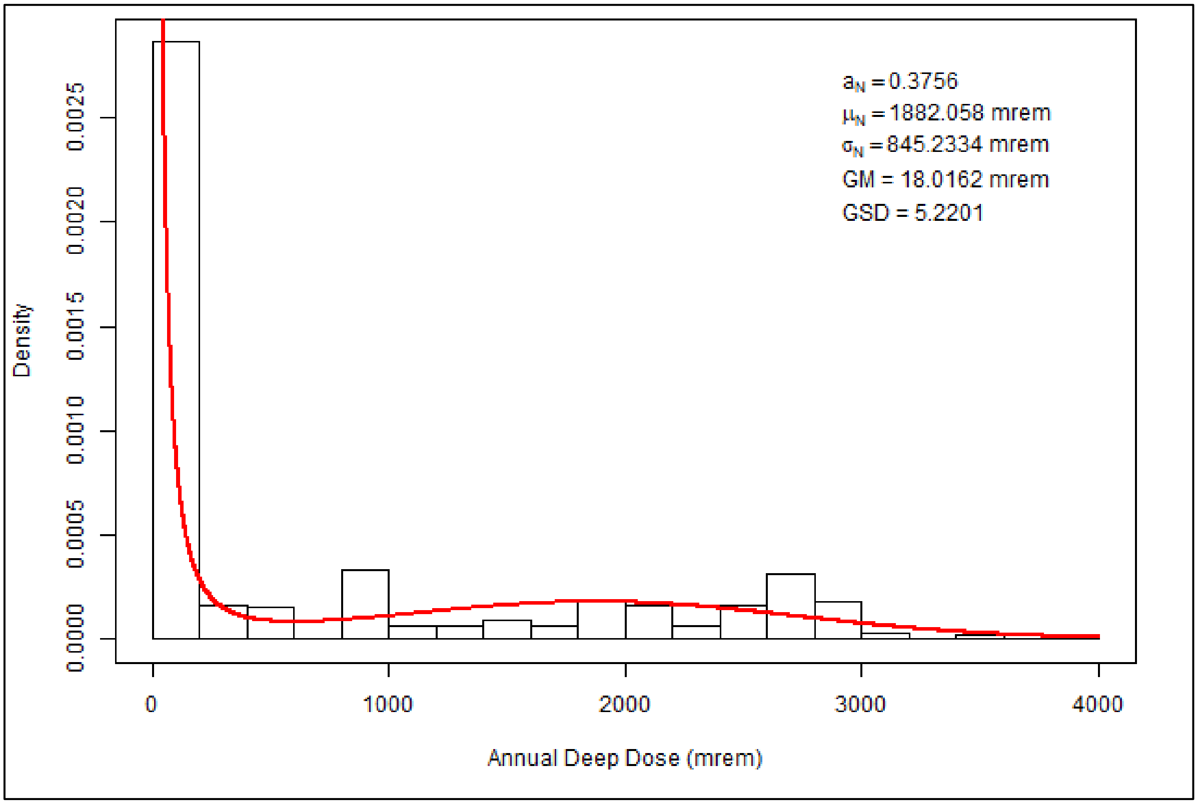

Figure 9-6. Histogram of Y-12 annual deep doses with fitted normal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

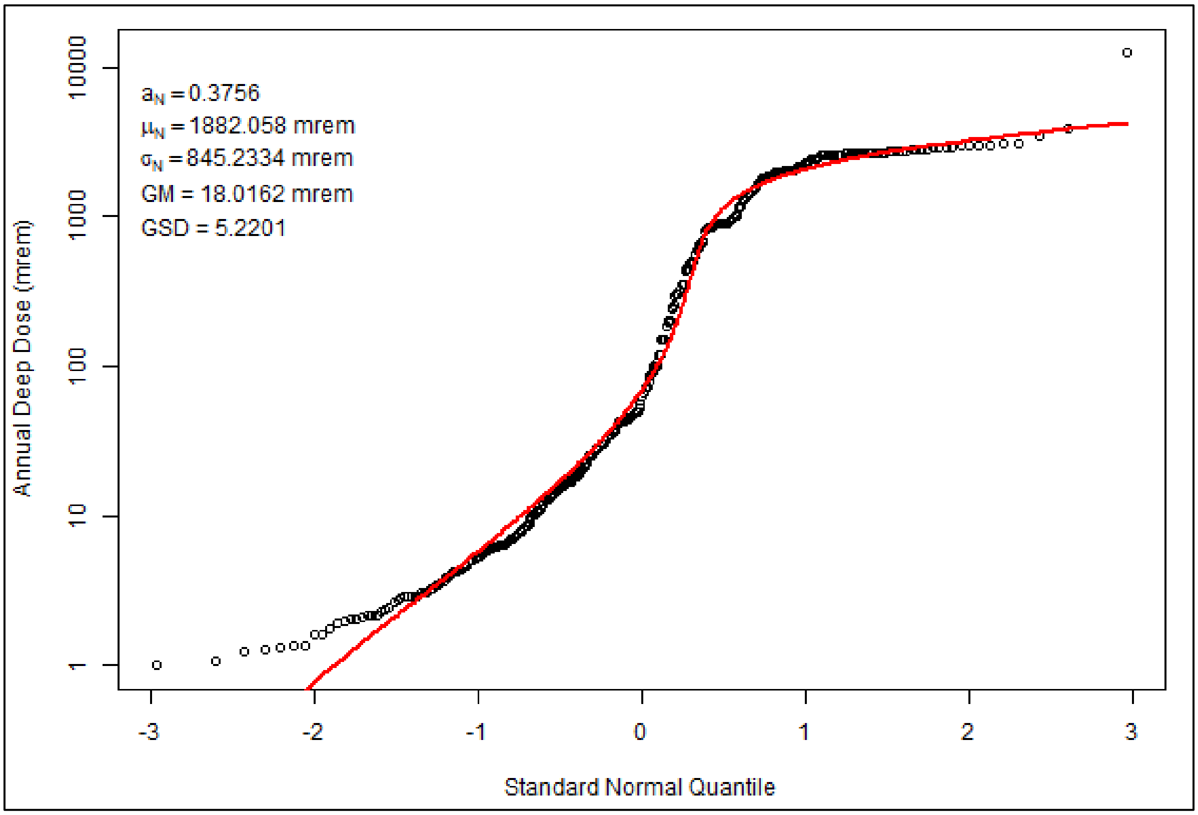

Figure 9-7. Probability plot of Y-12 annual deep doses with fitted normal-lognormal mixture. Attachment A contains an extended description.

Table 9-1. Summary of various measures of variability for Y-12 annual deep doses.

Measure of variability	Empirical	Fitted lognormal	Fitted normal	Fitted normal-lognormal mixture
Range	(1.0033, 12336)	(0, ∞)	(−∞, ∞)	(−∞, ∞)
IQR	1,453.9923	543.173	1,263.5785	1,522.8452
Variance	1.4641E6	5.0367E9	8.7739E5	1.0826E6
Standard deviation	1,209.9901	70,969.9075	936.6922	1,040.4655
$Q_{0.95}/Q_{0.05}$	1,291.1174	4,592.9484	−2.8986	1,751.2661
$Q_{0.95}/Q_{0.5}$	43.8423	67.7713	3.0534	41.1387

10.0 ENVIRONMENTAL DATA EXAMPLE

10.1 DESCRIPTION OF DATA

The environmental data in this section are thoron results from the Feed Materials Production Center (Fernald). This analysis considers 1,393 total thoron results with 196 results less than or equal to zero and 328 censored results. This means that 524 (or 38%) of the total thoron results are either less than or equal to zero or censored. The Fernald thoron data cover several orders of magnitude, so plotting the histogram on a logarithmic scale is helpful, but it does not allow for the results less than or equal to zero to be included. Using the censored results at face value in the histogram is also inappropriate. Therefore, Figure 10-1 is a histogram of the Fernald thoron results that are positive and uncensored.

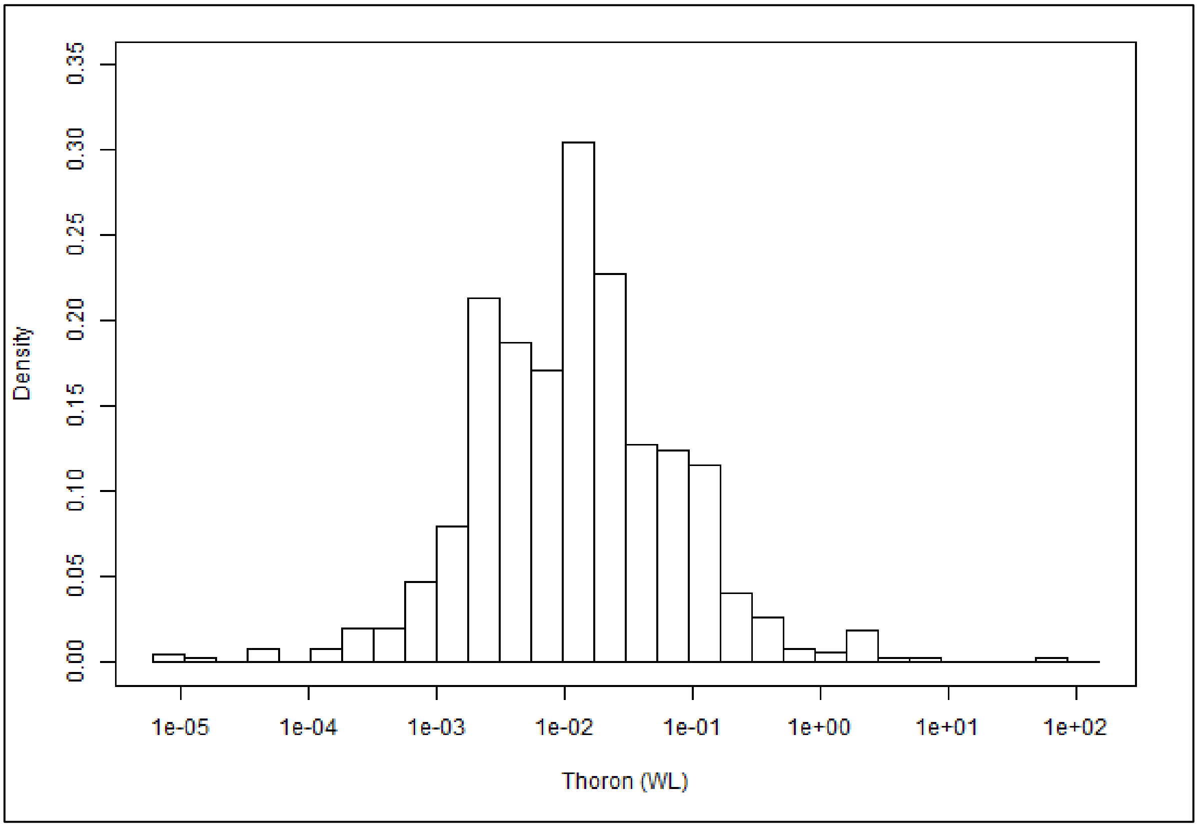

Figure 10-1. Histogram of Fernald thoron results that are positive and uncensored. Attachment A contains an extended description.

10.2 LOGNORMAL FIT

A lognormal distribution was fit to the thoron results. The lognormal fit appropriately handles the results that are less than or equal to zero or censored, but the histogram excludes them. Therefore, there is no histogram with the overlaid PDF from the lognormal fit. Figure 10-2 is a lognormal probability plot. The lognormal distribution fits reasonably well, except for the lower portion. Note that the GSD of the lognormal fit is 11.286, which is fairly large.

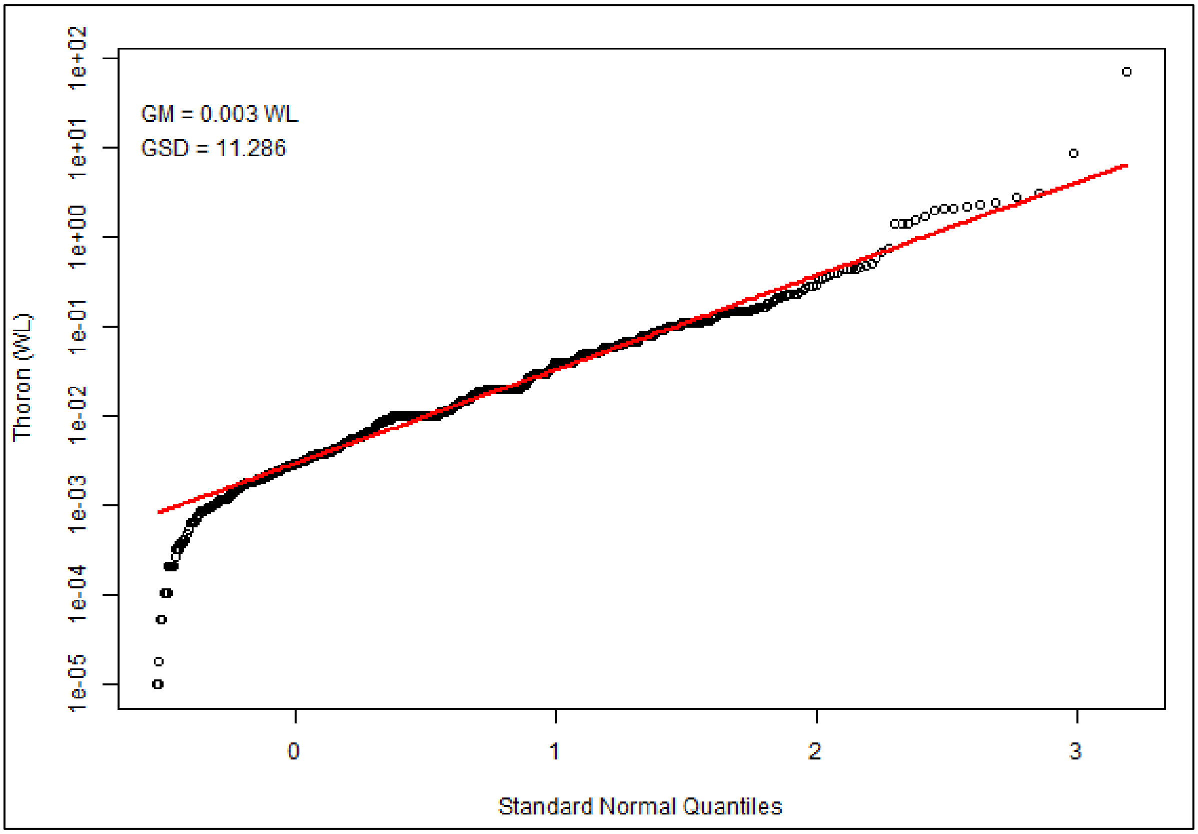

Figure 10-2. Lognormal probability plot of Fernald thoron results with fitted lognormal. Attachment A contains an extended description.

10.3 STRATIFICATION

In the dataset from which the Fernald thoron data were extracted, detector type is included: track etch or non-track etch. Figure 10-3 is a three-panel plot that includes the histogram from Figure 10-1 as the upper-left panel. The other two panels are histograms (on the same *x*-scale) of the two detector types. Figure 10-4 contains side-by-side boxplots for each detector type.

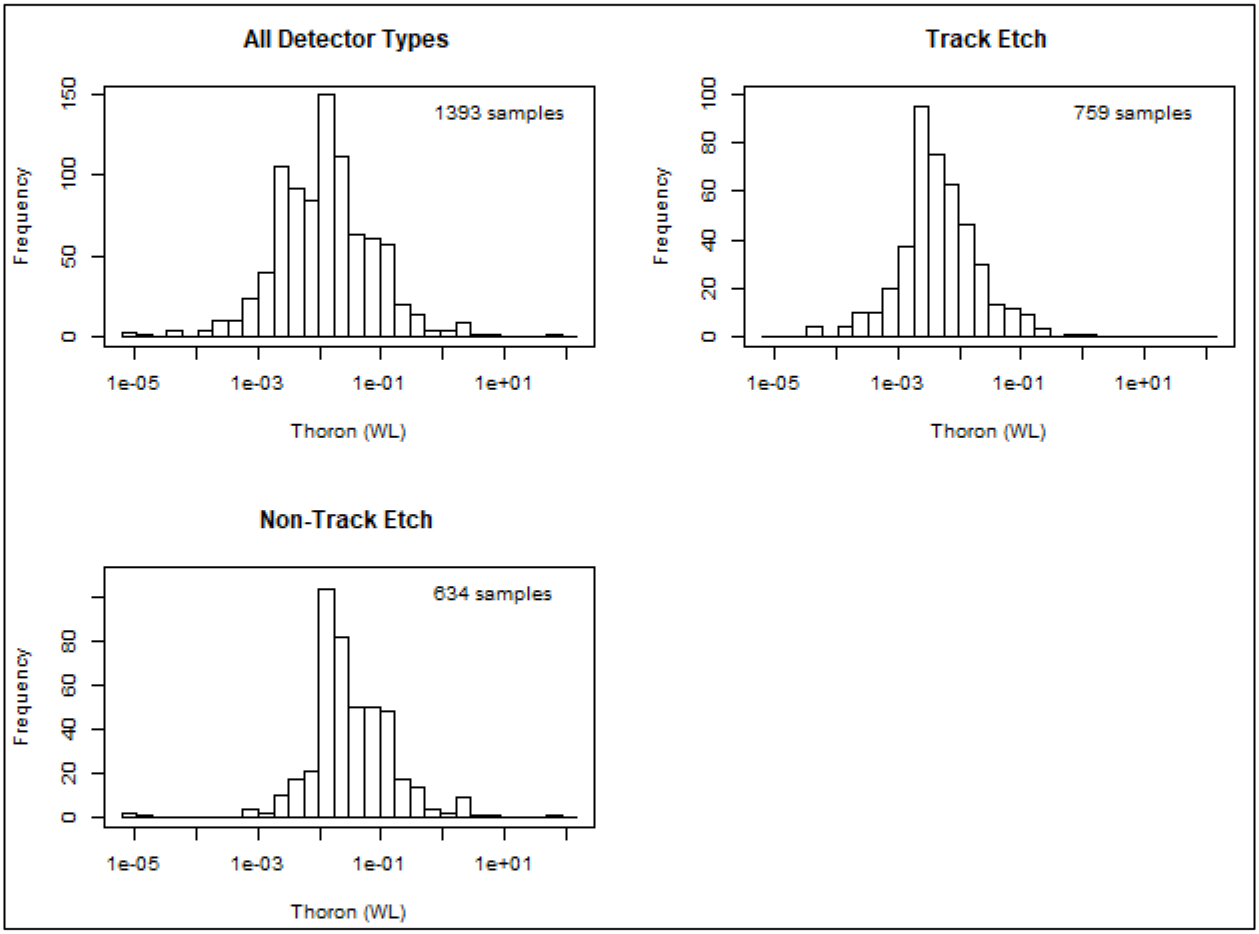

Figure 10-3. Three-panel plot of histograms of all detector types, track etch, and non-track etch. Attachment A contains an extended description.

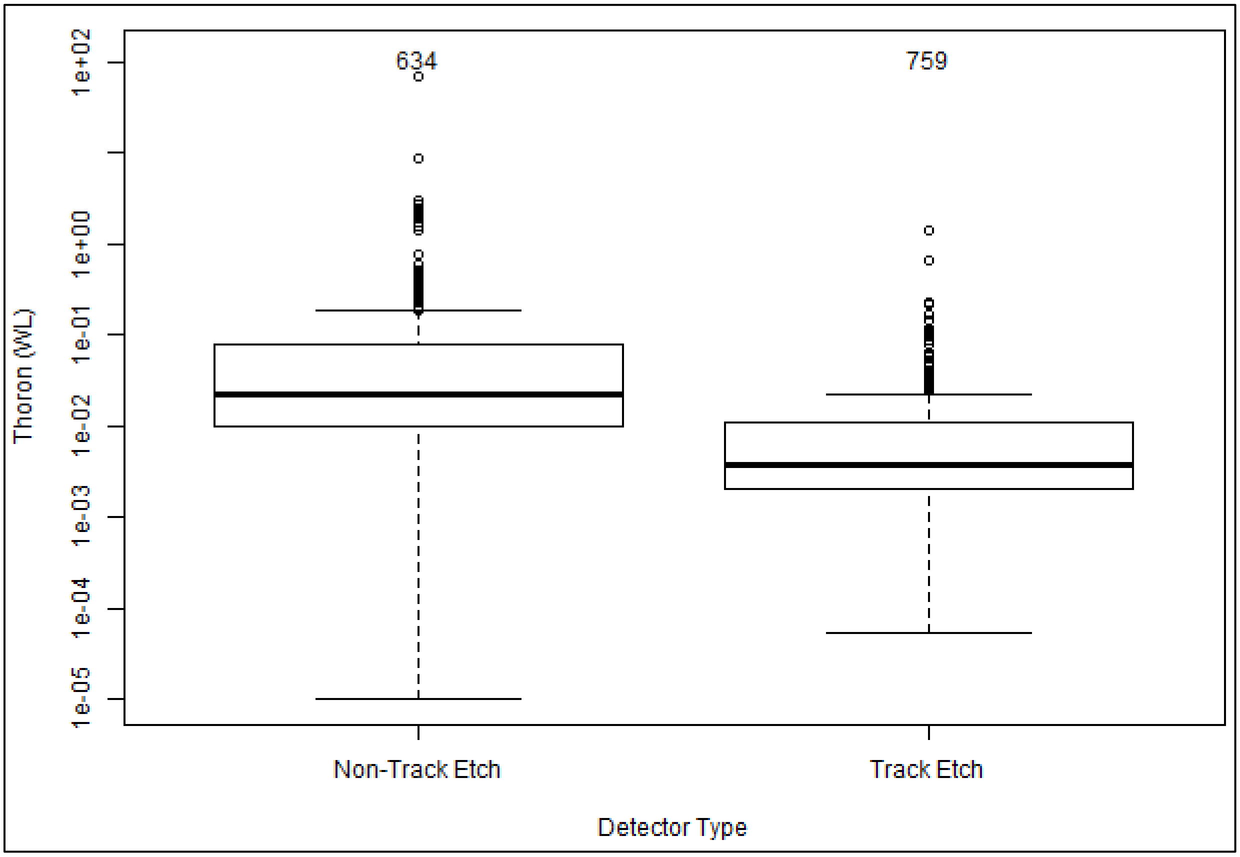

Figure 10-4. Side-by-side boxplots for each detector type: track etch and non-track etch. Attachment A contains an extended description.

From Figures 10-3 and 10-4, the empirical distribution of track etch looks different from non-track etch. Therefore, the subsequent analysis examines track-etch samples separately from non-track etch samples.

10.3.1 Track-Etch Samples

A lognormal distribution was fit to the track-etch samples. Figure 10-5 is a lognormal probability plot. The lognormal fit is reasonably good. Note that the GSD of the track-etch lognormal fit is 7.3112.

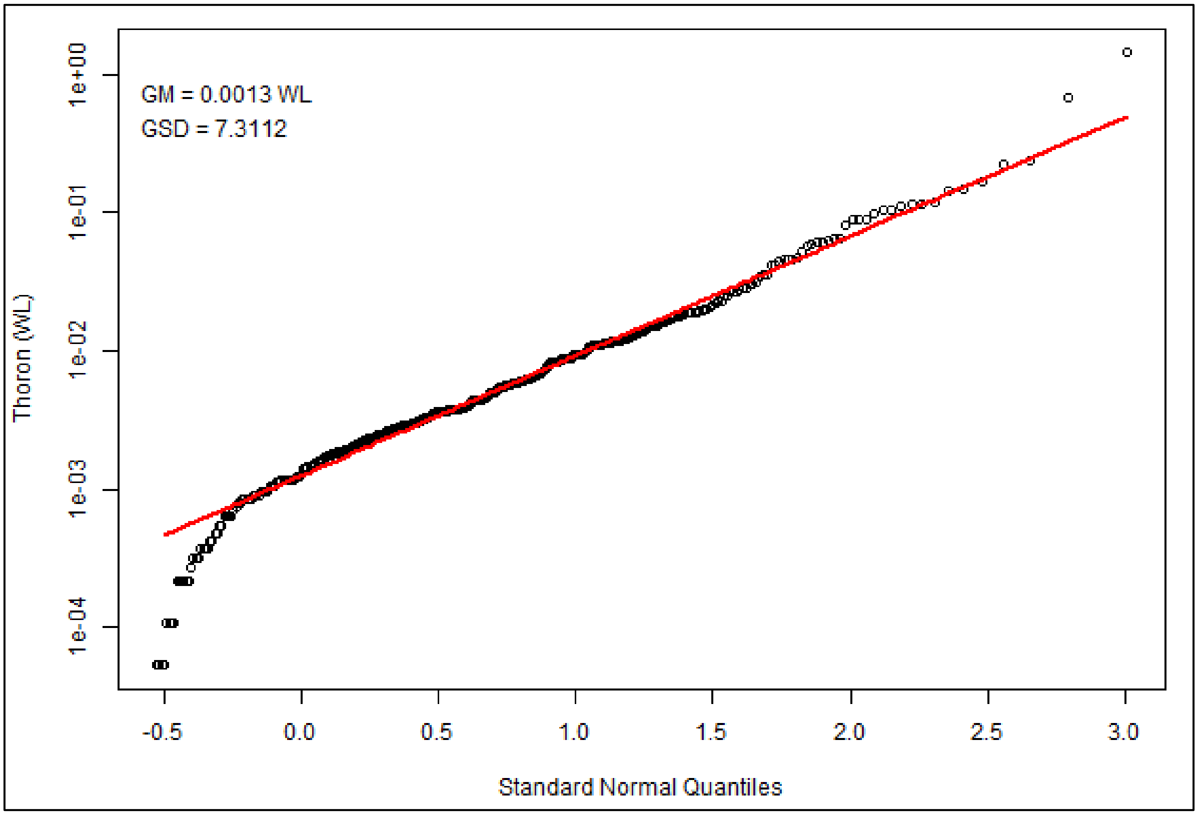

Figure 10-5. Lognormal probability plot of track-etch Fernald thoron results with fitted lognormal. Attachment A contains an extended description.

10.3.2 Non-Track Etch Samples

A lognormal distribution was fit to the non-track etch samples. Figure 10-6 is a lognormal probability plot. The lognormal fit is reasonably good. Note that the GSD of the non-track etch lognormal fit is 9.295.

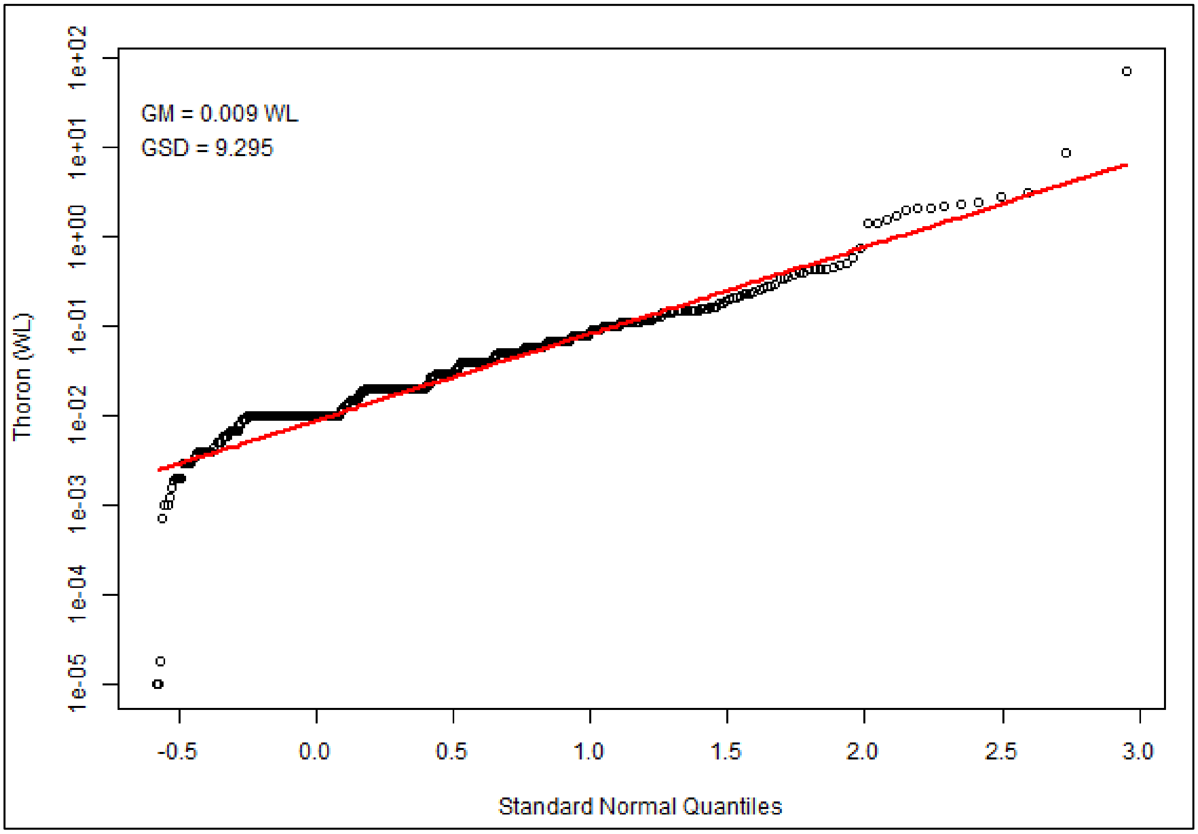

Figure 10-6. Lognormal probability plot of non-track etch Fernald thoron results with fitted lognormal. Attachment A contains an extended description.

If these stratified data were actually being analyzed for use, these stratified groups could be fit with other distributions (perhaps fitting mixture distributions to the stratified data) to improve the fits. The detector type was one of the available stratification variables, so other stratification variables could be considered. Because these examples are only for illustrative purposes, the analysis of the stratified groups ends here.

10.4 LOGNORMAL-LOGNORMAL MIXTURE FIT

To fit a lognormal-lognormal mixture distribution, all the data must be positive and uncensored. For this analysis, the thoron results that were less than or equal to zero or censored were imputed (single imputation). Details of the imputation can be found in ORAUT-RPRT-0096, *Multiple Imputation Applied to Bioassay Co-Exposure Models* [ORAUT 2021, pp. 7–8], and the support files [ORAUT 2024].

A lognormal-lognormal mixture distribution was fit to the Fernald thoron results. After imputation, all the thoron results are positive and uncensored, so a histogram with the mixture fit is appropriate. Figure 10-7 shows the histogram with overlaid PDF from a lognormal-lognormal mixture fit. Figure 10-8 is a probability plot. Based on both Figures 10-7 and 10-8, the lognormal-lognormal mixture distribution fits well. The fitted parameters indicate that about 20% of the mixture distribution is from the first lognormal and 80% from the second lognormal. Note that the GSDs of the lognormal portions of the mixture fit are 3.6434 and 8.6631, which is an improvement over the GSD of 11.286 from Figure 10-2.

Table 10-1 summarizes various measures of variability of the data and the distributions fit to the data. The empirical column can be compared to the lognormal column, and the empirical after imputation column can be compared to the lognormal-lognormal mixture column.

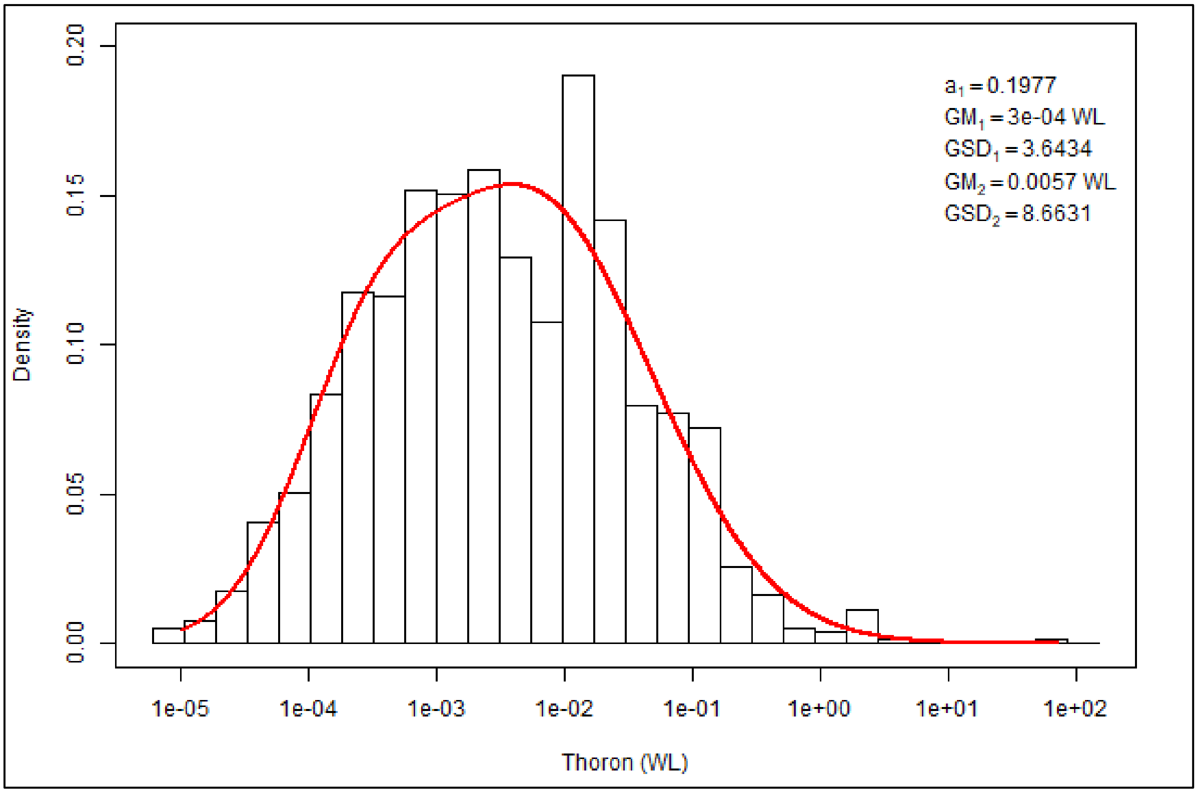

Figure 10-7. Histogram of Fernald thoron results with fitted lognormal-lognormal mixture PDF overlaid. Attachment A contains an extended description.

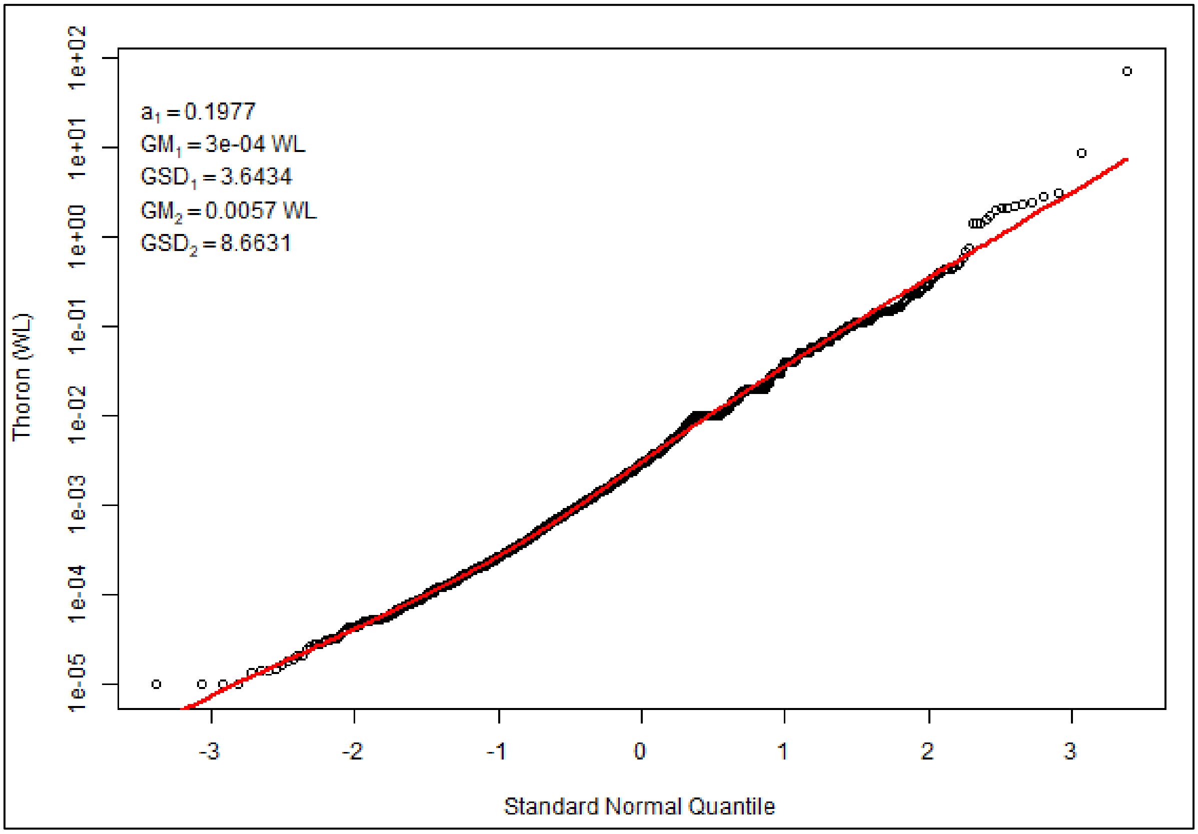

Figure 10-8. Probability plot of Fernald thoron results with fitted lognormal-lognormal mixture. Attachment A contains an extended description.

Table 10-1. Summary of various measures of variability for Fernald thoron results.

Measure of variability	Empirical	Fitted lognormal	Empirical after imputation	Fitted lognormal-lognormal mixture
Range	(−0.0024, 71)	(0, ∞)	(1E−5, 71)	(0, ∞)
IQR	0.0146	0.0146	0.0156	0.0160
Variance	3.7123	1.1013	3.7124	0.2913
Standard deviation	1.9267	1.0494	1.9268	0.5398
$Q_{0.95}/Q_{0.05}$	Undefined[a]	2,901.0366	1,847.0681	2,044.2310
$Q_{0.95}/Q_{0.5}$	32.1037	53.8613	47.0089	51.8189

a. The empirical 5th percentile is zero, so this ratio divides by zero.

11.0 CONCLUSIONS

This report defined variability, discussed the difference between variability and uncertainty, and explored causes of large variability. The lognormal distribution was presented, and the normal-lognormal mixture and lognormal-lognormal mixture distributions were presented as more complex alternatives when fitting health physics data. Three examples were analyzed to illustrate the use of stratification and mixture distributions, as they pertain to characterizing variability.

With historical data, reducing variability is impossible but it is important to adequately characterize and quantify the variability. For a dataset, if (1) outliers have been addressed, (2) stratification has been considered, and (3) the possibility of exploring various fitted distributions has been considered, then whatever variability is calculated from the data or from the fit is an acceptable characterization and quantification of the variability. Recall the question from Section 1.0: “how big a GSD is too big?” If outliers have been addressed, stratification has been explored, and the lognormal distribution is deemed to be an acceptable distribution, then the GSD of the lognormal fit is acceptable even if it could be perceived as excessively large. If an analysis adequately characterizes and quantifies the variability in the data, there is no numerical cutoff for how big a GSD is too big.

REFERENCES

Ellison SLR, Barwick VJ, Farrant TJD [2009]. Practical statistics for the analytical scientist: a bench guide. 2nd ed. Cambridge, UK: Royal Society of Chemistry. [SRDB Ref ID: 183169]

Farcomeni A, Geraci M [2024]. Quantile ratio regression. Stat Comput *34*(94):1–15. [SRDB Ref ID: 206836]

Gilbert RO [1987]. Statistical methods for environmental pollution monitoring. New York: Van Nostrand Reinhold Company. [SRDB Ref ID: 193116]

Huber PJ [1981]. Robust statistics. New York: John Wiley and Sons. [SRDB Ref ID: 206837]

Limpert E, Stahel WA, Abbt M [2001]. Log-normal distributions across the sciences: keys and clues. BioScience *51*(5):341–352. [SRDB Ref ID: 206838]

NIOSH [2019]. Comparison of several percentile definitions. Cincinnati, OH: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health. December. [SRDB Ref ID: 199956]

ORAUT [2014]. Analysis of stratified coworker datasets. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-RPRT-0053 Rev. 02, October 8. [SRDB Ref ID: 136245]

ORAUT [2015]. External dose coworker methodology. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-RPRT-0071 Rev. 00, July 2. [SRDB Ref ID: 145135]

ORAUT [2018]. Applications of regression in external dose reconstruction. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-RPRT-0087 Rev. 00, March 13. [SRDB Ref ID: 170100]

ORAUT [2021]. Multiple imputation applied to bioassay co-exposure models. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-RPRT-0096 Rev. 01, June 11. [SRDB Ref ID: 187045]

ORAUT [2024]. ORAUT-RPRT-0110 support files. Zip file. Oak Ridge, TN: Oak Ridge Associated Universities Team. December 4. [SRDB Ref ID: 206835]

Venables WN, Ripley BD [2002]. Modern applied statistics with S. 4th ed. New York: Springer. [SRDB Ref ID: 206840]

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES

Figure 3-1

The x-axis is log-scaled, labeled “X” and ranges from 5e-1 to 1e3. The y-axis is log-scaled, labeled “Y” and ranges from 5 to 2000. There are 30 points in a fairly linear pattern from about (5e-1, 5) to (5e1, 2000), and there is one point at about (1e3, 500).

Figure 3-2

The basic plot is the same as Figure 3-1. This plot has a dashed blue line from about (5e-1, 20) to (1e2, 1000) and text that says, “lm slope = 0.7351, lm intercept = 3.4489.” The red solid and purple dot-dashed lines are almost on top of each other from about (5e-1, 10) to (5e1, 1000). Text says, “rlm slope = 0.9625, rlm intercept = 3.0499.” Text says, “50th slope = 0.9819, 50th intercept = 2.9671.”

Figure 4-1

The x-axis is labeled “Simulated Data” and ranges from 0 to 50. The y-axis is labeled “Density” and ranges from 0 to 0.14. The histogram is made up of around 25 bars; the first bar is second-highest (about 0.1), the second bar is highest (about 0.11), and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, “GM = 5, GSD = 3.”

Figure 4-2

The x-axis is labeled “Simulated Data” and ranges from 0 to 50. The y-axis is labeled “Density” and ranges from 0 to 0.14. The histogram is made up of around 25 bars; the first bar is second-highest (about 0.1), the second bar is highest (about 0.11), and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, “GM = 5.0001, GSD = 3.”

Figure 4-3

The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is log-scaled, labeled “Empirical Quantile” and ranges from 0.2 to 200. The 1000 points are in a tight diagonal pattern from the bottom left to the top right corner. There is a straight red line going through the points. The text in the upper left corner says, “GM = 5.0001, GSD = 3.”

Figure 5-1

The x-axis is labeled “Simulated Data” and ranges from -2 to 15. The y-axis is labeled “Density” and ranges from 0 to 0.4. The histogram is made up of around 18 bars; the first bar is about 0.01, the second bar is about 0.23, the third bar is highest at about 0.28, and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, “a sub N = 0.5, mu sub N = 0, sigma sub N = 0.5, GM = 5, GSD = 3.”

Figure 5-2

The x-axis is labeled “Simulated Data” and ranges from -2 to 15. The y-axis is labeled “Density” and ranges from 0 to 0.4. The histogram is made up of around 18 bars; the first bar is about 0.01, the second bar is about 0.23, the third bar is highest at about 0.28, and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that starts at x of 0 and goes straight up, off the plot. It comes back down at about (0.5, 0.4) and looks like an exponential curve that tails off with the histogram bars. The text in the upper right corner reads, “GM = 0.786, GSD = 9.5515.”

Figure 5-3

The x-axis is labeled “Standard Normal Quantile” and ranges from -1 to 3. The y-axis is log-scaled, labeled “Empirical Quantile” and ranges from 1e-3 to 1e2. The 1000 points are in curved, monotonic increasing, mostly concave down pattern from the bottom left to the top right corner. There is a

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

straight red line going through (-1, 1e-1) and (2.5, 1e2). The text in the upper left corner says, "GM = 0.786, GSD = 9.5515."

Figure 5-4
The x-axis is labeled "Simulated Data" and ranges from -2 to 15. The y-axis is labeled "Density" and ranges from 0 to 0.4. The histogram is made up of around 18 bars; the first bar is about 0.01, the second bar is about 0.23, the third bar is highest at about 0.28, and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that is bell-shaped, starts at (-2, 0.05), peaks at (3.2, 0.1), and approaches a density of 0 by about x=15. The text in the upper right corner reads, "mu sub N = 3.238, sigma sub N = 4.4354."

Figure 5-5
The x-axis is labeled "Standard Normal Quantile" and ranges from -3 to 3. The y-axis is labeled "Empirical Quantile" and ranges from 0 to 150. The 1000 points are in curved, monotonic increasing, mostly concave up pattern from the bottom left to the top right corner. There is a straight red line going through (-2.5, -5) and (3, 15). The text in the upper left corner says, "mu sub N = 3.238, sigma sub N = 4.4354."

Figure 5-6
The x-axis is labeled "Simulated Data" and ranges from -2 to 15. The y-axis is labeled "Density" and ranges from 0 to 0.4. The histogram is made up of around 18 bars; the first bar is about 0.01, the second bar is about 0.23, the third bar is highest at about 0.28, and the remaining bars get shorter from left to right. There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, "a sub N = 0.5001, mu sub N = 1e-4, sigma sub N = 0.5001, GM = 5.0025, GSD = 2.9957."

Figure 5-7
The x-axis is labeled "Standard Normal Quantile" and ranges from -3 to 3. The y-axis is labeled "Empirical Quantile" and ranges from 0 to 150. The 1000 points are in curved, monotonic increasing, mostly concave up pattern from the bottom left to the top right corner. There is a red curve that goes through the points. The text in the upper left corner says, "a sub N = 0.5001, mu sub N = 1e-4, sigma sub N = 0.5001, GM = 5.0025, GSD = 2.9957."

Figure 6-1
The x-axis is labeled "Simulated Data" and ranges from 0 to 30. The y-axis is labeled "Density" and ranges from 0 to 0.1. The histogram is made up of around 30 bars. The distribution of the histogram is bimodal, with one mode around 1 (height of 0.083) and the other around 12.5 (height of 0.027). There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, "a sub 1 = 0.25, GM sub 1 = 2, GSD sub 1 = 2, GM sub 2 = 20, GSD sub 2 = 2."

Figure 6-2
The x-axis is labeled "Simulated Data" and ranges from 0 to 30. The y-axis is labeled "Density" and ranges from 0 to 0.1. The histogram is made up of around 30 bars. The distribution of the histogram is bimodal, with one mode around 1 (height of 0.083) and the other around 12.5 (height of 0.027). There is a red curve plotted over the bars that starts at (0, 0), rises to about (2.5, 0.06), and drops off to around (30, 0.01). The text in the upper right corner reads, "GM = 11.3145, GSD = 3.3203."

Figure 6-3
The x-axis is labeled "Standard Normal Quantile" and ranges from -3 to 3. The y-axis is log-scaled, labeled "Empirical Quantile" and ranges from 0.2 to 200. The 1000 points are in a monotonic S-shape

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

from top right to bottom left. There is a straight red line that goes through (-3, 0.2) and (2, 200). The text in the upper left corner says, "GM = 11.3145, GSD = 3.3203."

Figure 6-4
The x-axis is labeled "Simulated Data" and ranges from 0 to 30. The y-axis is labeled "Density" and ranges from 0 to 0.1. The histogram is made up of around 30 bars. The distribution of the histogram is bimodal, with one mode around 1 (height of 0.083) and the other around 12.5 (height of 0.027). There is a red curve plotted over the bars that follows the bars very well. The text in the upper right corner reads, "a sub 1 = 0.2488, GM sub 1 = 1.991, GSD sub 1 = 1.9835, GM sub 2 = 19.9628, GSD sub 2 = 2."

Figure 6-5
The x-axis is labeled "Standard Normal Quantile" and ranges from -3 to 3. The y-axis is log-scaled, labeled "Empirical Quantile" and ranges from 0.2 to 200. The 1000 points are in a monotonic S-shape from top right to bottom left. There is a curved red line that follows the shape of the points. The text in the upper left corner says, "a sub 1 = 0.2488, GM sub 1 = 1.991, GSD sub 1 = 1.9835, GM sub 2 = 19.9628, GSD sub 2 = 2."

Figure 8-3
The x-axis is labeled "Standard Normal Quantile" and ranges from 0 to 3.5. The y-axis is log-scaled, labeled "Pu-239 in Urine (pCi/day)" and ranges from 1e-4 to 1e1. The points are monotonic. Most points from about (0.25, 2e-4) to (2, 2e-2) are linear and follow the red line well. Below that portion, the points trail down away from the line. Above that portion, the points are slightly above the line at first, and tail off further from the line on the right. The straight line goes from (0, 2e-4) to (3.5, 1e0). The text in the upper left corner says, "GM = 2e-4 pCi/day, GSD = 12.709."

Figure 8-5
The x-axis is labeled "Standard Normal Quantile" and ranges from -3 to 3. The y-axis is labeled "Pu-239 in Urine (pCi/day)" and ranges from 0 to 20. The points are monotonic, relatively close to 0 pCi/day from -3 to around 2. Then the points tail up toward the highest point around (3.3, 20). The straight red line is fairly flat at around zero. The text in the upper left corner says, "mu sub N = 5e-4 pCi/day, sigma sub N = 0.0018 pCi/day."

Figure 8-6
This is a 4-panel plot, consisting of 4 histograms. The x-axes are all the same. The upper left plot is basically Figure 8-1, except the y-axis is "Frequency" and ranges from 0 to 450. The title of this panel is "All Schedule Types" and the text on the plot says, "2152 samples." The upper right plot is titled "Prompt Action" and the text on the plot says, "158 samples." The y-axis is labeled "Frequency" and ranges from 0 to 20. The bars are mostly between -0.005 and 0.02, but some range out to 0.1. The lower left plot is titled "Routine" and the text on the plot says, "1563 samples." The y-axis is labeled "Frequency" and ranges from 0 to 400. The bars are mostly between -0.005 and 0.005, but a few range out to 0.03. The lower right plot is titled "Special" and the text on the plot says, "431 samples." The y-axis is labeled "Frequency" and ranges from 0 to 100. The bars are mostly between -0.005 and 0.005, but a few range out to 0.025.

Figure 8-7
This plot contains 3 boxplots. The x-axis is labeled "Schedule Type" and includes (from left to right): "Prompt Action", "Routine", and "Special." The number of samples is 158, 1563, and 431, respectively. The y-axis is labeled "Pu-239 in Urine (pCi/day)" and ranges from 0 to 0.1. The prompt action box spans 0 to 0.02, with a line at 0.005. The whiskers extend to -0.005 and 0.05. There are open circles

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

from 0.058 to about 0.1. The routine and special boxplots are very similar. The boxes span 0 to 0.002, and the line in the middle is barely visible around 0.001. The whiskers extend to -0.005 and 0.005, with open circles from 0.005 to around 0.03.

Figure 8-8
The x-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from -0.005 to 0.1. The y-axis is labeled “Density” and ranges from 0 to 450. The bars are mostly between -0.005 and 0.005, but a few range out to 0.03. The red curve starts at (0, 450), dropping quickly, and almost to zero by 0.01. The text in the upper right corner says, “GM = 3e-4 pCi/day, GSD = 6.3608.”

Figure 8-9
The x-axis is labeled “Standard Normal Quantile” and ranges from 0 to 3.5. The y-axis is log-scaled, labeled “Pu-239 in Urine (pCi/day)” and ranges from 1e-4 to 1e1. The points are monotonic. Most points from about (0.25, 2e-4) to (2.7, 2e-2) are linear and follow the red line well. Below that portion, the points trail down away from the line. Above that portion, the points tail off upward from the line. The straight line goes from (0, 2e-4) to (3.5, 1e-1). The text in the upper left corner says, “GM = 3e-4 pCi/day, GSD = 6.3608.”

Figure 8-10
The x-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from -0.005 to 0.1. The y-axis is labeled “Density” and ranges from 0 to 450. The bars are mostly between -0.005 and 0.005, but a few range out to 0.03. The plot has a bell-shaped red curve starting at (-0.005, 0), rising to (0,250), and almost to zero by 0.005. The text in the upper right corner says, “mu sub N = 3e-4 pCi/day, sigma sub N = 0.0015 pCi/day.”

Figure 8-11
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from 0 to 20. The points are monotonic, relatively close to 0 pCi/day from -3 to around 2.75. Then the points tail up toward the highest point around (3.3, 20). The straight red line is fairly flat at around zero. The text in the upper left corner says, “mu sub N = 3e-4 pCi/day, sigma sub N = 0.0015 pCi/day.”

Figure 8-12
The x-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from -0.005 to 0.1. The y-axis is labeled “Density” and ranges from 0 to 250. The bars are mostly between -0.005 and 0.02, but some range out to 0.1. The red curve starts at (0, 250), dropping fairly quickly, and almost to zero by 0.04. The text in the upper right corner says, “GM = 0.0093 pCi/day, GSD = 10.081.”

Figure 8-13
The x-axis is labeled “Standard Normal Quantile” and ranges from -0.75 to 2.75. The y-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from 5e-4 to 5e0. The points are in a curved, monotonic increasing, mostly concave down pattern from the bottom left to the top right corner. There is a straight red line going through (-0.75, 1e-3) and (2.75, 5). The text in the upper left corner says, “GM = 0.0093 pCi/day, GSD = 10.081.”

Figure 8-14
The basic plot is the same as Figure 8-1. This plot has a fairly bell-shaped red curve starting at (-0.005, 0), rising to (0,300), and almost to zero by 0.02. The text in the upper right corner says, “a sub N = 0.8288, mu sub N = -2e-4 pCi/day, sigma sub N = 0.0012 pCi/day, GM = 0.0073 pCi/day, GSD = 5.7699.”

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

Figure 8-15
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is labeled “Pu-239 in Urine (pCi/day)” and ranges from 0 to 1. The points are monotonic, relatively close to 0 pCi/day from -3 to around 2. Then the points tail up toward the highest point around (3, 1). The curved red line follows the points fairly well. The text in the upper left corner says, “a sub N = 0.8288, mu sub N = -2e-4 pCi/day, sigma sub N = 0.0012 pCi/day, GM = 0.0073 pCi/day, GSD = 5.7699.”

Figure 9-1
The x-axis is labeled “Annual Deep Dose (mrem)” and ranges from 0 to 4000. The y-axis is labeled “Density” and ranges from 0 to 0.003. The histogram consists of 20 bars. The highest bar is the first bar with density of about 0.003. The other bars have heights ranging from 0 to 0.0004, with perhaps another mode in the tail.

Figure 9-3
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is log-scaled, labeled “Annual Deep Dose (mrem)” and ranges from 1 to 10000. The points are monotonic and in a severe S-shape from top right to bottom left. The straight line goes from (-2, 1) to (2, 10000). The text in the upper left corner says, “GM = 99.5403 mrem, GSD = 12.9777.”

Figure 9-4
The basic plot is the same as Figure 9-1. This plot has a bell-shaped red curve starting at (0, 0.0003), reaching its maximum at (800, 0.0005), and almost to zero by 3000 mrem. The text in the upper right corner says, “mu sub N = 750.3194 mrem, sigma sub N = 936.6922 mrem.”

Figure 9-5
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is labeled “Annual Deep Dose (mrem)” and ranges from 0 to 12000. The points are monotonic and in a severe S-shape from middle right to bottom left, with an “outlier” at (3, 12000). The straight line goes from (-1.3, 0) to (3, 3500). The text in the upper left corner says, “mu sub N = 750.3194 mrem, sigma sub N = 936.6922 mrem.”

Figure 9-6
The basic plot is the same as Figure 9-1. This plot has a red curve starting at (0, 0.003), dropping to (500, 0.0001), rising back up to (2000, 0.00025), and almost to zero by 4000 mrem. The text in the upper right corner says, “a sub N = 0.3756, mu sub N = 1882.058 mrem, sigma sub N = 845.2334 mrem, GM = 18.0162 mrem, GSD = 5.2201.”

Figure 9-7
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is log-scaled, labeled “Annual Deep Dose (mrem)” and ranges from 1 to 10000. The points are monotonic and in a severe S-shape from top right to bottom left. The curved red line follows the points fairly well except for a few points in the lower left corner. The text in the upper left corner says, “a sub N = 0.3756, mu sub N = 1882.058 mrem, sigma sub N = 845.2334 mrem, GM = 18.0162 mrem, GSD = 5.2201.”

Figure 10-1
The x-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The y-axis is labeled “Density” and ranges from 0 to 0.35. The histogram consists of 30 bars. The bars are in almost a bell-shape, but there appear to be two modes: one at 2e-3 (height of 0.22) and one at 1e-2 (height of 0.3).

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

Figure 10-2
The x-axis is labeled “Standard Normal Quantile” and ranges from -0.5 to 3. The y-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The points are monotonic and most of them follow the red line, except for several of them in the bottom left corner of the plot. The straight red line goes from (-0.5, 1e-3) to (3, 1e1). The text in the upper left corner says, “GM = 0.003 WL, GSD = 11.286.”

Figure 10-3
This is a 3-panel plot, consisting of 3 histograms. The x-axes are all the same. The upper left plot is basically Figure 10-1, except the y-axis is “Frequency” and ranges from 0 to 150. The title of this panel is “All Detector Types” and the text on the plot says, “1393 samples.” The upper right plot is titled “Track Etch” and the text on the plot says, “759 samples.” The y-axis is labeled “Frequency” and ranges from 0 to 100. The bars are mostly between 1e-5 and 1. The lower left plot is titled “Non-Track Etch” and the text on the plot says, “634 samples.” The y-axis is labeled “Frequency” and ranges from 0 to 100. The bars are mostly between 1e-3 and 1e1.

Figure 10-4
This plot contains 2 boxplots. The x-axis is labeled “Detector Type” and includes (from left to right): “Non-Track Etch”, and “Track Etch.” The number of samples is 634 and 759, respectively. The y-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The non-track etch box spans 1e-2 to 1e-1, with a line in the middle. The whiskers extend to 1e-5 and 2e-1. There are open circles from 2e-1 to about 1e2. The track etch box spans 1e-3 to 1e-2, with a line in the middle. The whiskers extend to 1e-4 and 2e-2, with open circles from 2e-2 to around 1.

Figure 10-5
The x-axis is labeled “Standard Normal Quantile” and ranges from -0.5 to 3. The y-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-4 to 1. The points are monotonic and most of them follow the red line, except for several of them in the bottom left corner of the plot. The straight red line goes from (-0.5, 5e-4) to (3, 2e-1). The text in the upper left corner says, “GM = 0.0013 WL, GSD = 7.3112.”

Figure 10-6
The x-axis is labeled “Standard Normal Quantile” and ranges from -0.5 to 3. The y-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The points are monotonic and most of them follow the red line, except for several of them in the bottom left corner of the plot. The straight red line goes from (-0.5, 2e-3) to (3, 1e1). The text in the upper left corner says, “GM = 0.009 WL, GSD = 9.295.”

Figure 10-7
The x-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The y-axis is labeled “Density” and ranges from 0 to 0.2. The histogram consists of 30 bars. The bars are in almost a bell-shape, but there appear to be two modes: one at 1e-3 (height of 0.15) and one at 1e-2 (height of 0.2). The red curve appears to be fairly unimodal, but not quite bell-shaped, because there are a couple of “lumps” near the top of the curve. The text in the upper left corner says, “a sub 1 = 0.1977, GM sub 1 = 3e-4 WL, GSD sub 1 = 3.6434, GM sub 2 = 0.0057 WL, GSD sub 2 = 8.6631.”

Figure 10-8
The x-axis is labeled “Standard Normal Quantile” and ranges from -3 to 3. The y-axis is log-scaled, labeled “Thoron (WL)” and ranges from 1e-5 to 1e2. The points are monotonic and almost linear, except for a little wiggle near the middle of the plot. The red curve appears to be a straight line but is

ATTACHMENT A
EXTENDED DESCRIPTIONS OF FIGURES (continued)

not, and it follows the points very well. The text in the upper left corner says, “a sub 1 = 0.1977, GM sub 1 = 3e-4 WL, GSD sub 1 = 3.6434, GM sub 2 = 0.0057 WL, GSD sub 2 = 8.6631.”