Journal of Occupational and Environmental Hygiene, 7: 233–244
ISSN: 1545-9624 print / 1545-9632 online
Copyright © 2010 JOEH, LLC
DOI: 10.1080/15459621003609713

An Accurate Substitution Method for Analyzing Censored Data

Gary H. Ganser[1] and Paul Hewett[2]

[1]Department of Mathematics, West Virginia University, Morgantown, West Virginia
[2]Exposure Assessment Solutions, Inc.; Morgantown, West Virginia

When analyzing censored datasets, where one or more measurements are below the limit of detection (LOD), the maximum likelihood estimation (MLE) method is often considered the gold standard for estimating the GM and GSD of the underlying exposure profile. A new and relatively simple substitution method, called β-substitution, is presented and compared with the MLE method and the common substitution methods (LOD/2 and LOD/$\sqrt{2}$ substitution) when analyzing a left-censored dataset with either single or multiple censoring points. A computer program was used to generate censored exposure datasets for various combinations of true geometric standard deviation (1.2 to 4), percent censoring (1% to 50%), and sample size (5 to 19 and 20 to 100). Each method was used to estimate four parameters of the lognormal distribution: (1) the geometric mean, GM; (2) geometric standard deviation, GSD; (3) 95th percentile, and (4) Mean for the censored datasets. When estimating the GM and GSD, the bias and root mean square error (rMSE) for the β-substitution method closely matched those for the MLE method, differing by only a small amount, which decreased with increasing sample size. When estimating the Mean and 95th percentile the β-substitution method bias results closely matched or bettered those for the MLE method. In addition, the overall imprecision, as indicated by the rMSE, was similar to that of the MLE method when estimating the GM, GSD, 95th percentile, and Mean. The bias for the common substitution methods was highly variable, depending strongly on the range of GSD values. The β-substitution method produced results comparable to the MLE method and is considerably easier to calculate, making it an attractive alternative. In terms of bias it is clearly superior to the commonly used LOD/2 and LOD/$\sqrt{2}$ substitution methods. The rMSE results for the two substitution methods were often comparable to rMSE results for the MLE method, but the substitution methods were often considerably biased.

Address correspondence to: Paul Hewett, 1270 Kings Road, Morgantown, WV 26508; e-mail: phewett.080641@oesh.com.

INTRODUCTION

As exposure limits decrease and exposure controls improve, an increasingly frequent occurrence is the left-censored dataset. This is a dataset where one or more measurements is less than the (field) limit of detection (LOD) (calculated using the laboratory LOD and the sample volume; also called the minimum detectable concentration). Various methods are in common use for estimating, from such a dataset, the parameters of the exposure profile (i.e., geometric mean and geometric standard deviation), and statistics such as the 95th percentile or the Mean exposure. These include maximum likelihood estimation (MLE),[1−4] log-probit regression (LPR),[4−6] and the LOD/2 and LOD/$\sqrt{2}$ substitution methods.[1,4,7]

Gilliom and Helsel[8] and She[9] used computer simulations to compare the performance of the LOD/2 substitution method with more sophisticated methods and found that the substitution methods tend to produce biased results when estimating the mean of a lognormal distribution, and, for this reason, should be avoided. In contrast, Hornung and Reed[1] suggested whenever the percent censored is less than 50%, that the LOD/2 substitution method could be used in epidemiologic studies to estimate the mean exposure. More recently, Hewett and Ganser[4] recommended against the indiscriminate use of the LOD/2 and LOD/$\sqrt{2}$ substitution methods, but recognized that there are scenarios where it might be expedient to use a judiciously selected substitution method.

Although it has been clearly demonstrated in the literature that the substitution methods are biased and that superior methods are available, both the LOD/2 and LOD/$\sqrt{2}$ substitution methods continue to be used, particularly in the occupational and environmental epidemiologic literature where it is expedient, for example, to use a substitution method when statistically analyzing a large exposure database or when constructing a job exposure matrix.

Hornung and Reed[1]showed that for very large sample sizes, the two common substitution methods, which substitute each LOD with the LOD multiplied by either 1/2 or 1/$\sqrt{2}$, are biased when estimating the exposure profile geometric mean (GM) and geometric standard deviation (GSD), and this bias can be positive or negative, depending on the true GSD and the true fraction or percent censored. (In this article, the LOD is determined from the percent censored. For example, the LOD for a distribution that is 10% censored is that concentration where the cumulative probability function is 0.10.)

In a previous article,[4] we reproduced the results of Hornung and Reed but also looked at smaller sample sizes. We found it interesting that their tables and our own results (for smaller, more typical sample sizes) suggest that for specific combinations of true GSD, true percent censored, and sample size, the bias for the LOD/2, and LOD/$\sqrt{2}$ substitution methods can approach zero. This led us to investigate whether it is possible to devise a substitution method where a factor can be calculated (from the uncensored data in the dataset) that results on average in a near zero bias. The resulting method, which we call β-substitution, turned out to be both easy to implement and surprisingly accurate, comparable to the MLE method in terms of both bias and overall imprecision (expressed as the root mean square error or rMSE). This article compares the β-substitution method with the MLE method, as well as the LOD/2 and LOD/$\sqrt{2}$ substitution methods, when applied to a left-censored dataset with both single and multiple censoring points, or LOD values.

BACKGROUND

Censored data analysis (CDA) methods tend to fall into one of four categories:

1. substitution methods, i.e., LOD/2 and LOD/$\sqrt{2}$;[1,7]
2. log-probit regression (LPR) methods;[5,6]
3. maximum likelihood estimation (MLE) methods;[1–3] and
4. various non-parametric methods.

Methods have been recommended that were deemed useful whenever there are multiple detection limits or it is suspected that the true distribution is not well described by the lognormal distribution. These include the non-parametric Kaplan-Meier survival analysis method and the "multiple-detection limit" method (a variation on the LPR method).

Interested readers are referred to Helsel[10] for a discussion of these methods, as well as variations on the MLE and LPR methods. More recently, Hewett and Ganser[4] used computer simulation to compare the most commonly used CDA methods. They concluded that the standard MLE method comes closest to being an omnibus method and should be preferred when estimating the mean or upper percentiles, such as the 95th percentile.

In this article, we focus on the bias and rMSE of the proposed β-substitution method, comparing them with the bias and rMSE of the MLE method and the common substitution methods when analyzing both simple and complex censored datasets. A simple censored dataset contains measurements censored at a single LOD, or two or more LODs, but all at the low end of the dataset. In contrast, a complex-censored dataset contains measurements censored at two or more LODs with uncensored measurements scattered in between.

MLE METHOD

The MLE method is considered the "gold standard" method, provided the data are well described by a lognormal distribution. The maximum likelihood (sample) GM and GSD are those values that maximize the likelihood function:

$$LF = \prod_{i=k+1}^{n} pdf(\ln x_i | lnGM, lnGSD) \cdot \prod_{j=1}^{k} cdf(\ln x_j | lnGM, lnGSD) \qquad \textbf{(1)}$$

where n = sample size (including both censored and uncensored data), k = number of censored data, the x_i values are the detects and the x_j values are the non-detects, pdf refers to the normal probability density function, and cdf refers to the normal cumulative density function. Since there is no close-form solution to this equation, the maximum likelihood sample GM and GSD are generally determined using the Newton-Raphson non-linear function maximization technique. (Finkelstein and Verma[3] recently showed how the solver function of a spreadsheet can be used to find the optimal solution.)

After calculating the sample GM and GSD, a compliance statistic, such as the sample 95th percentile, can be estimated using the standard equation:

$$\hat{X}_{0.95} = \exp(\ln(gm) + 1.645 \cdot \ln(gsd)) \qquad \textbf{(2)}$$

where gm = sample GM and gsd = sample GSD.[11]

If a dataset is completely uncensored, the preferred estimator for the mean of lognormally distributed data is the minimum variance unbiased estimator (MVUE)[6,12] (although there is usually little practical difference between the simple arithmetic mean and the MVUE). The MVUE equation corrects for the transformation bias inherent in moving from the log scale to the concentration scale. We found through experimentation that this formula can also be used when estimating the mean of a censored dataset. The MVUE equation requires three inputs:

1. ln(gm),
2. ln(gsd), and
3. the sample size.

The MLE method provides estimates of the GM and GSD, but what sample size should be used in the MVUE equation? We tried both the full sample size (n) and the number of uncensored data (n-k). We found again through experimentation that between the two, the full sample size n in the MVUE equation results in less bias. Therefore, for the MLE method, we estimated the mean using the MVUE equation and n as the sample size.

Mathematically, the MLE method can be applied to datasets as small as n = 3, provided there are at least two uncensored data. However, as we found during our initial simulations, if n is less than 10, the estimates of GM and GSD can be severely biased, particularly when the true percent censored is large.

LOD/2 and LOD/$\sqrt{2}$ Substitution Methods

The common substitution methods are the easiest to implement and have the advantage of automatically accommodating multiple LODs. Each LOD is simply replaced with an appropriate value—LOD/2 or LOD/$\sqrt{2}$—followed by a conventional statistical analysis of the revised dataset. After estimating the sample GM and GSD, compliance statistics such as the sample 95th percentile can be estimated using Eq. 2. The Mean is estimated using the standard formula for the simple arithmetic mean. The minimum sample size for substitution methods is $n = 2$, with at least one uncensored datum.

Hornung and Reed[1] used computer simulation to show that for very large sample sizes ($n = 100{,}000$) the bias for the substitution methods can be substantially positive or negative for the GM and GSD, depending on the true GSD and the true percent censored. They offered advice on when each of the two substitution methods could be used to produce estimates of the GM or GSD. They also suggested using LOD/2 substitution when estimating the mean of the lognormal distribution (as in an epidemiologic study). However, they did not determine bias for the Mean, nor did they estimate the bias for the GM and GSD when the sample size is small, say, less than 100, or for upper percentiles of the distribution, such as the 95th percentile.

Later, El-Shaarwai and Esterby[13] derived formulae for directly calculating the large sample bias for the Mean when using substitution methods, given known values for the GM, GSD, and percent censored. However, their formulae cannot be used to determine the bias for a specific small sample size or the rMSE at any sample size, and they did not address bias or rMSE when estimating upper percentiles. Specifically, for this article, their formulae do not lead to an improved substitution method because, in their view, the correct factor to replace 1/2 or 1/$\sqrt{2}$ depends on parameters whose values are not known (i.e., the true GM and GSD). They suggested an iterative procedure using their formulae but concluded that this new method, in practice, would be functionally identical to an existing algorithm for the MLE method and therefore did not represent an improvement.

β-SUBSTITUTION METHOD ALGORITHM

Hornung and Reed[1] found that for specific true GSDs and percent censored, each of the two substitution methods will produce nearly unbiased estimates of the true GM and GSD. That is, multiplying each LOD by 1/2 or 1/$\sqrt{2}$ is suited for specific combinations of true GSD and true percent censored (they used a single sample size of 100,000). This led to an inquiry into whether a factor, which we call β, can be calculated for a specific dataset that when used in place of 1/2 or 1/$\sqrt{2}$ will result in a near zero bias (and low rMSE).

The β-substitution method has as its foundation similar theoretical considerations as those presented by El-Shaarwai and Esterby[13] but with a different implementation (Appendix A). Estimates for the unknown parameter values are based on the number of uncensored and observed data, and no iteration is involved. The β-substitution method involves the calculation, from the uncensored data in the dataset, of a β factor for adjusting each LOD value. The β factor varies, depending on whether the GM or Mean is being estimated. After estimating, the GM and Mean the GSD is estimated, after which the sample 95th percentile (or other percentiles) can be calculated. The β-substitution algorithm involves a modest number of steps, each of which can be implemented using a spreadsheet or the programming language of most statistical analysis programs.

Step 1: Create a detects array (i.e., an array of the uncensored data). Let n = total sample size and k = number of measurements $<$ LOD.

Step 2: Calculate input and intermediate values:

$$\bar{y} = \frac{1}{n-k}\sum_{i=1}^{n-k} y_i \tag{3}$$

where $y_i = \ln(x_i)$ and x_i is the *ith* detect

$$z = \Phi^{-1}\left[\frac{k}{n}\right] \tag{4}$$

$$f(z) = \frac{pdf(z,0,1)}{1 - cdf(z,0,1)} \tag{5}$$

$$\hat{s}_y = \frac{\bar{y} - \ln(LOD)}{f(z) - z} \tag{6}$$

$$f(\hat{s}_y, z) = \frac{1 - cdf(z - \frac{\hat{s}_y}{n}, 0, 1)}{1 - cdf(z,0,1)} \tag{7}$$

The function Φ^{-1} refers to the inverse of the Z-distribution; i.e., the Z-value that corresponds to k/n. The statements pdf[z,0,1] and cdf[z,0,1] refer to the probability density function for the unit normal (i.e., Z-value) distribution and the cumulative density function for the unit normal distribution, respectively. The denominator can also be calculated as (n–k)/n. The s_y quantity is an initial estimate of the ln(gsd).

Step 3: Calculate β_{MEAN}:

$$\beta_{\text{MEAN}} = \frac{n}{k}\cdot cdf(z - \hat{s}_y, 0, 1)\cdot \exp\left[-\hat{s}_y\cdot z + \frac{(\hat{s}_y)^2}{2}\right] \tag{8}$$

Step 4: Substitute each non-detect with $\beta_{\text{MEAN}}\cdot$ LOD, and calculate the sample mean, i.e., the simple arithmetic mean, using a combined array of detects and substituted measurements.

Step 5: Calculate β_{GM}:

$$\beta_{\text{GM}} = \exp\left[\frac{-(n-k)\cdot n}{k}\cdot \ln(f(\hat{s}_y, z)) - \hat{s}_y\cdot z - \frac{n-k}{2kn}\cdot(\hat{s}_y)^2\right] \tag{9}$$

Step 6: Substitute each non-detect with $\beta_{\text{GM}}\cdot$ LOD, and calculate the sample GM using the combined array of detects and substituted measurements.

Step 7: Recalculate s_y and then calculate the sample GSD:

$$s_y = \sqrt{\frac{2n}{n-1}\cdot\ln\left(\frac{mean}{gm}\right)} \tag{10}$$

$$gsd = \exp(s_y) \tag{11}$$

Note: If (mean/gm) ≤ 1, which happens occasionally when the sample size is small and the detects are close in value to the LOD, then let $s_y = 0$ and gsd = 1.

Step 8: Calculate the sample 95th percentile:

$$\hat{X}_{0.95} = \exp\left[\ln(gm) - \frac{s_y^2}{2n} + 1.645\cdot s_y\right] \tag{12}$$

(For a different percentile, the z-value of 1.645 should be replaced by the appropriate value.)

Like the MLE method, the minimum sample size is n = 3 with at least two uncensored data, but in practice, the sample size should be 5 or greater for a percent censored up to 50%.

In our experience, most censored datasets appear to be of the simple-censored variety; that is, a single laboratory is used per year resulting in a single LOD or two or more LODs, but all at the low end. A complex-censored dataset can occur whenever an investigator combines data from several studies where different laboratories were used, or combines data from exposure groups that were collected across some broad span of time during which several laboratories were utilized. In the situation where there are multiple censoring points, an average field LOD can be calculated for use in the above equations:

$$L\bar{O}D = \exp\left[\frac{1}{\sum m_i}\cdot\sum(m_i\cdot\ln(LOD_i))\right] \tag{13}$$

where m_i is the number of measurements censored at the *ith* LOD.

METHODS

To estimate the bias and overall accuracy for each method, we developed a computer program to generate censored datasets that then were analyzed using the MLE method, β-substitution, and the LOD/2 and LOD/$\sqrt{2}$ substitution methods. For each artificial dataset, the estimates of the four parameters:

1. the sample GM,
2. GSD,
3. 95th percentile, and
4. Mean,

were compared with the true value. The average bias and the overall rMSE were calculated (using formulae presented later) after 100,000 datasets were created and analyzed.

After the sample GM and GSD were determined for each dataset, the sample 95th percentile was calculated using Eq. 2 (except for the β-substitution method, which has a slightly modified formula). For the MLE method, the Mean was estimated using the MVUE equation (Mulhausen and Damiano[(6)] or Bullock and Ignacio).[(11)] For the LOD/2, LOD/$\sqrt{2}$, and β-substitution methods, the Mean was estimated using the standard formula for the simple arithmetic mean.

To compare the methods we devised two simulations (Table I):

TABLE I. Parameters Used in the Computer Simulations

Simulation Parameter	Simulation 1, Scenario I: Single Lognormal Distribution Single LOD	Simulation 1, Scenario II: Contaminated Lognormal Distribution Multiple LODs	Simulation 2: Small Sample Size; Single Lognormal Distribution and Single LOD
	Min–Max	Min–Max	Min–Max
Sample Size	20–100	20–100	5–19
		Exposure Profile Distributions	
GM_1	1	1–3	1
GSD_1	(1.2–2) (2–3) (3–4) (1.2–4)	(1.2–2) (2–3) (3–4) (1.2–4)	(1.2–2) (2–3) (3–4) (1.2–4)
GM_2	—	1–3	—
GSD_2	—	(1.2–2) (2–3) (3–4) (1.2–4)	—
$Distribution_1$ %^{A}	—	0–100%	—
		Lab LOD as % of the Exposure Profile^B	
Lab_1 LOD %	1–50%	1–50%	1–50%
Lab_2 LOD %	—	1–50%	—
Lab_3 LOD %	—	1–50%	—

Note: Separate simulations were run using the each of the four ranges for the GSD.

^{A}Distribution$_2$% will be 1, the percentage for Distribution$_1$.

^{B}If three labs were used, each was used one-third of the time.

- Simulation 1: n ranged between 20 and 100, true percent censored ranged between 1% and 50%, and the true GSD ranged between 1.2 and 2, 2 and 3, 3 and 4, and 1.2 and 4
- Simulation 2: n ranged between 5 and 19, true percent censored ranged between 1% and 50%, and the true GSD ranged between 1.2 and 2, 2 and 3, 3 and 4, and 1.2 and 4

For Simulation 1 we devised two scenarios:

- Scenario I: a single lognormal distribution and a single LOD
- Scenario II: a contaminated lognormal distribution and three LODs

For each Simulation 1, Scenario, and method combination, we generated 100,000 artificial datasets from censored lognormal or censored contaminated lognormal distributions. The sample size for each dataset was randomly varied (using the uniform distribution) between 20 and 100 (inclusive). The percentage of the distribution that was censored was also randomly varied (using the uniform distribution), between 1% and 50% (inclusive). The field LOD (i.e., minimum detectable concentration) was then set at the concentration in the distribution corresponding to the percent censored.

In Scenario I, a single lognormal distribution was assumed as well as a single laboratory. The GSD for the distribution was randomly varied between 1.2 and 4 (inclusive) using the uniform distribution. The methods were also challenged with Scenario II — where the distribution was not truly lognormal (i.e., a contaminated lognormal distribution) as well as multiple LODs — due to a concern in the censored data literature (Helsel[(10)] for a review) that a valid CDA method should not only be fairly robust to departures from the lognormal model but also capable of handling multiple LODs.

In Scenario II, three laboratories with different field LODs were simulated, one each for approximately one-third of the samples. The field LOD for each lab was randomly generated as described above. In addition, the underlying distribution was contaminated. A contaminated (i.e., non-lognormal) distribution was created by combining two lognormal distributions. The GM and GSD for each distribution were randomly generated from uniform distributions where the minimum and maximum values were 1 and 3, and 1.2 and 4, respectively. The fraction that the first distribution contributed to the overall distribution was also randomly varied by generating a fraction from a uniform distribution where the minimum and maximum were 0 and 1. (The fraction contributed by the second distribution was one minus this value.)

For each artificial dataset, the program determined if the dataset was invalid, valid and censored, or valid and completely uncensored. A dataset was invalid if all n measurements were censored or there were too few detects (both the MLE and beta methods require at least two detects, whereas the LOD/2 and LOD/$\sqrt{2}$ substitution methods require only one detect). The fraction of invalid datasets for each method, GSD, sample size, and percent censored combination was tracked by the simulation program.

Since the minimum sample size in Simulation 1 was 20 and the maximum percent censored was 50%, the percentage of invalid datasets was negligible. (For Simulation 1 and both Scenarios I and II, the fraction of invalid datasets was zero. For Simulation 2 the typical fraction was less than 0.5%, as the smaller sample sizes increased the probability of an invalid or completely censored dataset.) The selected censored data analysis method was applied whenever the dataset was valid and censored (i.e., there is at least one LOD value). If all n measurements were uncensored, the dataset was statistically analyzed by the program using standard statistical methods (Mulhuasen and Damiano[(6)] or Bullock and Ignacio[(11)]).

Once the sample GM, GSD, 95th percentile, and Mean were calculated, the differences between the sample estimates and the true values are determined. For all 100,000 datasets the program calculated the average bias for each of the parameters:

$$Bias = (\bar{x} - \theta) \tag{14}$$

Where $\bar{x}$ is the mean of the 100,000 parameter estimates and θ is the true value. The program also calculated the rMSE, which is a combination of bias and precision:

$$rMSE = \sqrt{(\bar{x} - \theta)^2 + \frac{\sum (x - \bar{x})^2}{N - 1}} \tag{15}$$

where N = 100,000. The rMSE value can be considered a measure of overall accuracy or imprecision.

RESULTS AND DISCUSSION

Results are presented in Tables II–IV. The bias and rMSE for each of four parameters

1. GM,
2. GSD,
3. 95th percentile, and
4. Mean are given.

We will focus on the results for the 95th percentile and the Mean.

In Table II, where the "single lognormal distribution and single LOD scenario" was assumed for both the 95th percentile and Mean, the rMSE results for all of the methods were reasonably comparable, leaving it to the bias values to distinguish between the methods. The mean bias results for the MLE and β-substitution methods were comparable and closely aligned for all of the GSD ranges: the full GSD range (i.e., 1.2–4) as well as the three narrow GSD ranges. In contrast, the bias values for the substitution methods tended to vary substantially according to the GSD range, particularly for the 95th percentile.

In Table III, where a "contaminated lognormal distribution and multiple LOD scenario" was assumed, the bias and rMSE results for the MLE and β-substitution methods were again closely aligned while, in contrast, the bias values for the substitution methods tended to vary substantially.

TABLE II. Simulation 1, Scenario I: Single Lognormal Distribution and Single LOD Scenario

	Method			
GSD Range	**MLE**	**β-Substitution**	**LOD/2 Substitution**	**LOD/$\sqrt{2}$ Substitution**
		GM Bias		
1.2–2	0.1	−0.2	−9.0	−0.8
2–3	0.6	0.0	−1.5	7.7
3–4	1.3	0.1	0.0	14.8
1.2–4	0.8	0.0	−1.9	7.8
		GM rMSE		
1.2–2	7.4	7.5	12.8	7.5
2–3	14.0	14.2	13.4	16.1
3–4	24.7	19.4	19.3	25.6
1.2–4	14.9	14.8	15.5	18.5
		GSD Bias		
1.2–2	−0.3	0.4	12.3	0.0
2–3	−0.2	0.6	0.2	−9.5
3–4	0.0	0.1	−7.1	−15.6
1.2–4	−0.2	0.3	1.0	−9.0
		GSD rMSE		
1.2–2	6.2	6.5	14.8	6.2
2–3	12.1	12.6	9.8	13.9
3–4	16.8	17.4	14.7	20.5
1.2–4	12.8	13.2	13.2	15.2
		$X_{0.95}$ Bias		
1.2–2	−0.4	0.2	9.9	−0.7
2–3	0.3	0.2	−0.6	−8.3
3–4	1.6	−0.9	−5.9	1.3
1.2–4	0.6	−0.3	0.4	−7.8
		$X_{0.95}$ rMSE		
1.2–2	11.1	11.3	16.4	11.3
2–3	21.7	22.2	21.2	21.5
3–4	30.8	31.4	28.0	27.8
1.2–4	23.0	23.5	23.0	22.4
		Mean Bias		
1.2–2	−0.2	−0.1	−4.9	−1.1
2–3	−0.6	0.0	−1.2	1.3
3–4	−0.7	−0.2	0.0	1.3
1.2–4	−0.5	−0.1	−2.0	0.6
		Mean rMSE		
1.2–2	7.4	7.5	10.1	7.7
2–3	15.7	16.7	16.8	16.6
3–4	24.7	27.9	28.0	27.8
1.2–4	17.8	19.6	20.2	19.7

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD, percent censored, and sample size were randomly selected from ranges specified in Table I.

In the simulations above, the sample size was permitted to vary across a range of 20 to 100. Table IV lists the results when the methods were challenged with a "single lognormal distribution and single LOD scenario," and the sample size was restricted to the range of 5 to 19. Here the results were more interesting. Regarding rMSE, the β-substitution method was either comparable or superior to the MLE method. Regarding bias, the β-substitution method was clearly less biased than the MLE method. In contrast, while the rMSE results for the substitution methods were comparable to those of the MLE and β-substitution methods, the bias tended to vary substantially for the various GSD ranges. The

TABLE III. Simulation 1, Scenario II: Contaminated Lognormal Distribution and Multiple LOD Scenario

	Method			
GSD Range	**MLE**	**Substitution**	**LOD/2 Substitution**	**LOD/$\sqrt{2}$ Substitution**
GM Bias				
1.2–2	0.2	1.7	−8.4	−0.1
2–3	0.8	3.2	−1.1	7.9
3–4	1.4	4.7	5.2	14.8
1.2–4	1.1	4.0	−0.5	8.5
GM rMSE				
1.2–2	7.6	7.6	11.8	7.2
2–3	14.0	14.2	13.2	15.4
3–4	19.4	19.6	18.9	24.1
1.2–4	14.8	15.2	14.6	17.2
GSD Bias				
1.2–2	−0.6	−1.4	10.8	−1.2
2–3	−0.5	−2.2	−0.3	−9.8
3–4	−0.2	−3.0	−7.3	−15.7
1.2–4	−1.3	−1.8	1.2	−10.6
GSD rMSE				
1.2–2	7.4	7.7	13.1	6.1
2–3	12.1	13.7	9.3	13.1
3–4	16.3	17.4	13.6	19.2
1.2–4	13.8	15.9	11.6	15.0
$X_{0.95}$ Bias				
1.2–2	−0.4	−0.5	8.8	−1.5
2–3	0.3	−0.9	−0.9	−8.3
3–4	1.3	−1.2	−5.9	−12.1
1.2–4	−0.4	1.0	−1.5	−8.8
$X_{0.95}$ rMSE				
1.2–2	12.6	12.7	16.4	12.3
2–3	22.1	23.0	21.1	21.1
3–4	30.5	32.5	28.4	28.7
1.2–4	24.8	30.2	22.0	21.9
Mean Bias				
1.2–2	−0.3	0.8	−0.7	−4.4
2–3	−0.7	0.5	−1.0	1.4
3–4	−1.1	0.5	0.0	1.3
1.2–4	−2.2	0.8	−1.5	1.0
Mean rMSE				
1.2–2	7.6	7.6	9.6	7.8
2–3	15.9	16.9	17.3	17.1
3–4	24.6	28.6	28.4	28.7
1.2–4	18.1	21.7	22.0	21.9

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD, percent censored, and sample size were randomly selected from ranges specified in Table I.

results suggest that the β-substitution method is an attractive alternative to the MLE method whenever the sample size is small.

Regarding the GM and GSD, the bias and rMSE of the β-substitution method were similar to those of the MLE method for all of the GSD ranges. In contrast, the two substitution methods tended to produce biased estimates, with the bias varying substantially with the GSD range. However, the rMSE values tended to be comparable to those of the MLE method. Given similar rMSE values, the preferred method should be the one with the smallest bias.

To further compare the various methods, we contrived several additional simulations but only for the GSD range of 1.2 to 4 and only for the 95th percentile and Mean parameters.

TABLE IV. Simulation 2: Small Sample Size Scenario

	Method			
GSD Range	**MLE**	**β-Substitution**	**LOD/2 Substitution**	**LOD/$\sqrt{2}$ Substitution**
		GM Bias		
1.2–2	0.7	–0.4	–7.8	0.0
2–3	3.5	0.5	1.7	10.6
3–4	7.1	2.6	10.7	20.2
1.2–4	4.0	1.0	2.2	11.0
		GM rMSE		
1.2–2	16.1	16.5	19.3	15.5
2–3	31.9	31.4	30.4	31.9
3–4	46.3	44.9	44.9	48.8
1.2–4	34.6	33.9	34.2	35.8
		GSD Bias		
1.2–2	–1.2	2.1	11.8	–0.4
2–3	–0.8	3.3	0.2	–9.7
3–4	1.1	2.9	–6.5	–15.4
1.2–4	–0.2	2.5	1.0	–9.1
		GSD rMSE		
1.2–2	14.7	17.0	18.8	10.8
2–3	30.7	32.1	20.2	20.6
3–4	49.5	42.4	27.1	28.5
1.2–4	36.1	32.6	22.5	21.9
		$X_{0.95}$ Bias		
1.2–2	–1.2	1.3	10.7	0.1
2–3	2.5	0.3	5.1	–2.9
3–4	9.4	–2.1	6.1	–1.8
1.2–4	3.9	–0.7	6.9	–1.7
		$X_{0.95}$ rMSE		
1.2–2	25.0	25.3	29.9	24.3
2–3	53.5	47.1	52.9	49.1
3–4	87.6	66.2	82.4	74.1
1.2–4	61.5	49.2	59.9	54.0
		Mean Bias		
1.2–2	–0.8	–0.3	–5.0	–1.1
2–3	–1.9	–0.3	–1.2	1.4
3–4	–3.0	0.1	0.2	1.4
1.2–4	–2.0	–0.5	–2.0	0.6
		Mean rMSE		
1.2–2	15.8	16.3	18.4	16.3
2–3	33.6	35.7	36.2	36.1
3–4	53.4	62.6	63.0	60.3
1.2–4	38.1	42.1	43.1	42.2

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD, percent censored, and sample size were randomly selected from ranges specified in Table I.

First, we repeated Simulation 1, Scenario I but increased the "percent censored" to 50–80%. Next, we used the original range for "percent censored" of 1–50% but increased the range of the sample sizes to 100–1000. Last, we repeated Simulation 1, Scenario II (i.e., the contaminated lognormal distribution, multiple LOD scenario) but used a sample size range of 100–1000. The results (Tables V–VII) show that the bias and rMSE of β-substitution method consistently follows those of the MLE method. In contrast, both of the substitution methods had highly variable bias results and nearly always greater rMSE results. Again, the β-substitution method was clearly superior to the simple substitution methods and nearly

TABLE V. Simulation 1, Scenario I, Single Lognormal Distribution and Single LOD Scenario, with Lab LOD in the Range 50–80%

	Method			
GSD Range	MLE	β-Substitution	LOD/2 Substitution	LOD/$\sqrt{2}$ Substitution
		$X_{0.95}$ Bias		
1.2–4	1.1	–1.3	–19.8	–21.3
		$X_{0.95}$ rMSE		
1.2–4	25.4	23.7	21.2	27.7
		Mean Bias		
1.2–4	1.3	–0.2	–0.9	12.1
		Mean rMSE		
1.2–4	19.1	19.8	21.2	25.0

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD and sample size were randomly selected from ranges specified in Table I.

TABLE VII. Simulation 1, Scenario II: Contaminated Lognormal Distribution and Multiple LOD Scenario, with Sample Sizes in the Range 100–1000

	Method			
GSD Range	MLE	β-Substitution	LOD/2 Substitution	LOD/$\sqrt{2}$ Substitution
		$X_{0.95}$ Bias		
1.2–4	0.6	1.8	–21.6	–21.6
		$X_{0.95}$ rMSE		
1.2–4	10.1	12.6	23.9	23.6
		Mean Bias		
1.2–4	–2.4	0.9	0.2	12.3
		Mean rMSE		
1.2–4	8.3	8.5	10.2	16.0

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD and percent censored were randomly selected from ranges specified in Table I.

equal to the MLE method in terms of both bias and rMSE for the above simulations.

Although the MLE method and β-substitution method yield similar simulation results when estimating the GM, GSD, 95th percentile, and Mean, the sample statistics can and will vary for any particular dataset. This is because the computer simulations yield what could be called "long-run" results. In the long run, across numerous datasets, the results of the β-substitution method will be closer to the results of the MLE method than will the results of the LOD/2 and LOD/$\sqrt{2}$ substitution methods. For example, consider the following censored dataset (borrowed from Finkelstein and Verma[3]):

$$x = \{< 3, < 3, < 3, 3.06, 4.41, 7.23, 8.29, \times 9.52, 19.94, 20.25\}. \quad \textbf{(16)}$$

We estimated the GM, GSD, 95th percentile, and Mean using each of the methods used in this article. The results are shown in Table VIII. The results from the β-substitution method do not differ significantly from the MLE method results. The results for the LOD/$\sqrt{2}$ substitution method are clearly different from both the MLE and β-substitution results. However, results for the LOD/2 substitution method are similar to those of the β-substitution method. This has to do with the fact that for *this particular dataset* the β-substitution LOD adjustment factors for both the Mean and GM calculations are close to 0.5: $Mean = 0.5875$ and $GM = 0.4941$. Since the LOD adjustment factor for the LOD/2 substitution method is 0.5, it is no surprise that for this

TABLE VI. Simulation 1, Scenario I, Single Lognormal Distribution and Single LOD Scenario, with Sample Sizes in the Range 100–1000

	Method			
GSD Range	MLE	B-Substitution	LOD/2 Substitution	LOD/$\sqrt{2}$ Substitution
		$X_{0.95}$ Bias		
1.2–4	0.0	0.0	–0.8	–8.9
		$X_{0.95}$ rMSE		
1.2–4	7.9	8.7	12.2	14.1
		Mean Bias		
1.2–4	–0.1	0.0	–1.8	0.6
		Mean rMSE		
1.2–4	6.4	7.0	7.9	7.3

Notes: The bias and rMSE were calculated for 100,000 artificial datasets. For each dataset, the GSD and percent censored were randomly selected from ranges specified in Table I.

TABLE VIII. Parameter Estimates for the Finkelstein and Verma Dataset

	Estimate			
Parameter	MLE	β-Substitution	LOD/2 Substitution	LOD/$\sqrt{2}$ Substitution
GM	5.17	5.02	5.04	5.59
GSD	2.64	2.69	2.76	2.42
$X_{0.95}$	25.5	24.3	26.8	23.9
Mean	7.77	7.80	7.72	7.91

Note: Finkelstein and Verma.[3]

particular dataset the results of the LOD/2 and β-substitution methods are similar. But for the "long run" we recommend using the β-substitution method if the MLE method is not available.

In principle, the MLE method has the following advantages: it can be applied to virtually any sample size (but preferably 10 or more); it can be applied where the percent censored exceeds 50% (provided the sample size is large); it becomes increasingly more accurate as the sample size increases; and it handles complex datasets. On the negative side, the MLE method tends to be strongly biased whenever the sample size is small (e.g., $n < 10$), but its chief disadvantage is that it can be difficult to implement. However, simplified versions (e.g., the often cited Hald[14] or Cohen[15] methods) have been available for some time and continue to be used. Furthermore, the MLE method is often available in statistical programs and can be implemented using readily available statistical programming software.[10,16] Finkelstein and Verma[3] demonstrated the use of a spreadsheet solver function to find the optimal solution to the likelihood equation (Eq. 1); however, some manual data manipulation is still required.

The β-substitution method has the advantages of the MLE method, is nearly equal in performance, and is considerably easier to program using either common spreadsheet functions (as demonstrated in Appendix B) or the programming language built into most statistical packages. In addition, we suspect that β-substitution is more robust than the MLE method in that, in principle, it is not as susceptible to the effect of outliers due to the fact it does not use the estimate of the standard deviation of the uncensored data in the algorithm as does the MLE method.

There are other CDA methods besides the MLE method. There are several variations on the MLE method, several versions of the log-probit regression CDA method, and various non-parametric approaches to estimating the 95th quantile and mean. In a separate paper, Hewett and Ganser[4] showed that the MLE was remarkably robust when challenged with contaminated lognormal datasets, concluding that of all the methods tested, the standard MLE method comes closest to being an omnibus method: applicable to (nearly) all combinations of sample size, conformance with the lognormal distribution assumption, and percent censored, and whether the dataset is simple or complex. We compared the β-substitution results from this article with those for identical simulations and scenarios in Hewett and Ganser.[4] The β-substitution method, when compared with the several variations on the MLE and log-probit regression methods, was nearly always superior in terms of bias and was either comparable or superior in terms of rMSE.

In summary, the β-substitution method is easier to implement and has nearly the same performance of the MLE method: both the bias and rMSE results were comparable if not lower than those for the MLE method for the simulation and scenarios postulated. In contrast, the LOD/2 and LOD/$\sqrt{2}$ substitution methods were clearly biased when estimating the GM, GSD, 95th percentile, and Mean; the bias was often strongly positive or negative, depending on the true GSD, percent censored, and the sample size; and the bias did not approach zero as the sample size increased. (These findings regarding MLE method and the common substitution methods match those of Hornung and Reed[1].) Since the β-substitution method yields similar results and is considerably easier to implement, it is an attractive alternative to the MLE method, particularly for sample sizes less than 20 where it tended to have both smaller bias and rMSE. It is also clearly superior in all of the simulations to the two commonly used substitution methods.

REFERENCES

1. **Hornung, R.W., and L.D. Reed:** Estimation of average concentration in the presence of nondetectable values. *Appl. Occup. Environ. Hyg. 5*:46–51 (1990).
2. **Perkins, J.L., G.N. Cutter, and M.S. Cleveland:** Estimating the mean, variance, and confidence limits from censored (<Limit of Detection), lognormally distributed exposure data. *Am. Ind. Hyg. Assoc. J. 51*:416–419 (1990).
3. **Finkelstein, M.M., and D.K. Verma:** Exposure estimation in the presence of nondetectable values: Another look. *AIHAJ 62(2)*:195–198 (2001). Letters to the Editor, *AIHAJ 63(1)*:4 (2002).
4. **Hewett, P., and G.H. Ganser:** A comparison of several methods for analyzing censored data. *Ann. Occup. Hyg. 51*:611–632 (2007).
5. **Hawkins, N.C., S.K. Norwood, and J.C. Rock (eds.):** *A Strategy for Occupational Exposure Assessment*. Fairfax, Va.: AIHA, 1991.
6. **Mulhausen, J., and J. Damiano (eds.):** *A Strategy for Assessing and Managing Occupational Exposures,* Second Edition. Fairfax, Va.: AIHA, 1998.
7. **Glass, D.C., and C.N. Gray:** Estimating mean exposures from censored data: Exposure to benzene in the Australian petroleum industry. *Ann. Occup. Hyg. 45(4)*:275–282 (2001).
8. **Gilliom R.J., and D.R. Helsel:** Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques. *Water Resources Res. 22*:135–146 (1986).
9. **She, N.:** Analyzing censored water quality data using a non-parametric approach. *J. Am. Water Resources Assoc. 33*:615–624 (1997).
10. **Helsel, D.R.:** *Nondetects and Data Analysis*. New York: John Wiley & Sons, Inc., 2005.
11. **Bullock, W.H., and J.S. Ignacio (eds.):** *A Strategy for Assessing and Managing Occupational Exposures*, Third Edition. Fairfax, Va.: AIHA, 2006.
12. **Hewett, P.:** Industrial hygiene exposure assessment—Data analysis and interpretation. In *Handbook of Chemical Health and Safety*, R.J. Alaimo (ed.). Washington, D.C.: American Chemical Society, 2001.
13. **El-Shaarawi, A.H., and S.R. Esterby:** Replacement of censored observations by a constant: An evaluation. *Water Resources 26*:835–844 (1992).
14. **Hald, A.:** *Statistical Theory with Engineering Applications*. New York: John Wiley & Sons, Inc., 1952. pp. 144–151.
15. **Cohen, A.C.:** Tables for maximum likelihood estimates: Single truncated or single censored samples. *Technometrics 3*:535 (1961).
16. **Oak Ridge National Laboratory (ORNL):** *Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-detectable Values* by E.L. Frome and P.F. Wambach (ORNL/TM-2005/52). Oak Ridge, Tenn.: ORNL, 2005.

APPENDIX A

Derivation β_{GM} for a Single LOD

The method of moments is used to estimate the geometric mean by choice of the factor β_{GM}. Using the notation introduced earlier, the observed values y = ln(x) have the following distribution:

$$f(y) = \frac{pdf(y, lnGM, lnGSD)}{1 - cdf(\tilde{z}, 0, 1)} \qquad \textbf{(A1)}$$

where

$$\tilde{z} = \frac{lnLOD - lnGM}{lnGSD} \qquad \textbf{(A2)}$$

The sample estimate of the geometric mean is

$$\widehat{GM} = \left(\prod_{i=1}^{n-k} x_i\right)^{\frac{1}{n}} \cdot (\tilde{\beta}_{\mathrm{GM}} \cdot LOD)^{\frac{k}{n}} \qquad \textbf{(A3)}$$

Therefore, for fixed values of k and n,

$$E[\overset{\wedge}{GM}] = \left[\left(e^{\frac{lnGM}{n} + \frac{(ln GSD)^2}{2n^2}}\right) \cdot f(\ln GSD, \tilde{z})\right]^{n-k} \cdot (\tilde{\beta}_{\mathrm{GM}} \cdot LOD)^{\frac{k}{n}} \qquad \textbf{(A4)}$$

where

$$f(\ln GSD, \tilde{z}) = \frac{1 - cdf\left(\tilde{z} - \frac{\ln GSD}{n}, 0, 1\right)}{1 - cdf(\tilde{z}, 0, 1)} \qquad \textbf{(A5)}$$

This becomes $e^{\ln GM}$ by choosing

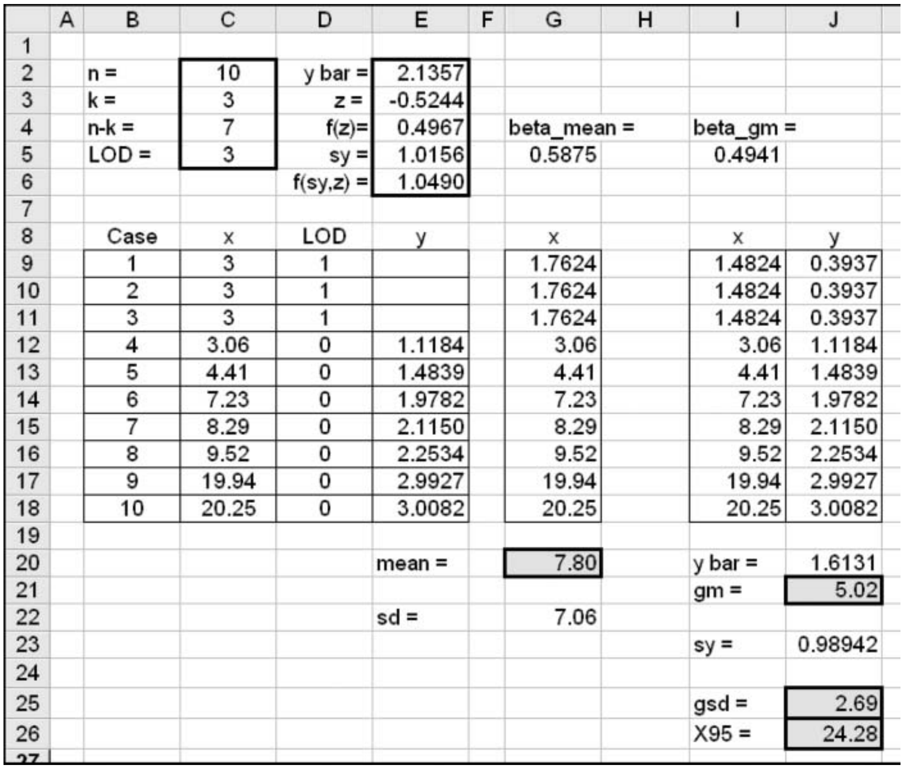
$$\tilde{\beta}_{\mathrm{GM}} = \exp\left[\frac{-(n-k) \cdot n}{k} \cdot \ln(f(lnGSD, \tilde{z})) - lnGSD \cdot \tilde{z} - \frac{n-k}{2kn} \cdot (lnGSD)^2\right] \qquad \textbf{(A6)}$$

Finally, $\tilde{z}$ is replaced with the estimate $z = \Phi^{-1}[k/n]$ and lnGSD with the estimate $\hat{s}_y$ to give β_{GM}. The estimate for $\hat{s}_y$ follows from the identity

$$E(\bar{y}) - \ln LOD = -\tilde{z} \cdot lnGSD + lnGSD \cdot f(\tilde{z}) \qquad \textbf{(A7)}$$

The derivation of β_{MEAN} is similar.

APPENDIX B



	A	B	C	D	E	F	G	H	I	J
1										
2		n =	10	y bar =	2.1357					
3		k =	3	z =	-0.5244					
4		n-k =	7	f(z)=	0.4967		beta_mean =		beta_gm =	
5		LOD =	3	sy =	1.0156		0.5875		0.4941	
6				f(sy,z) =	1.0490					
7										
8		Case	x	LOD	y		x		x	y
9		1	3	1			1.7624		1.4824	0.3937
10		2	3	1			1.7624		1.4824	0.3937
11		3	3	1			1.7624		1.4824	0.3937
12		4	3.06	0	1.1184		3.06		3.06	1.1184
13		5	4.41	0	1.4839		4.41		4.41	1.4839
14		6	7.23	0	1.9782		7.23		7.23	1.9782
15		7	8.29	0	2.1150		8.29		8.29	2.1150
16		8	9.52	0	2.2534		9.52		9.52	2.2534
17		9	19.94	0	2.9927		19.94		19.94	2.9927
18		10	20.25	0	3.0082		20.25		20.25	3.0082
19										
20					mean =		7.80		y bar =	1.6131
21									gm =	5.02
22					sd =		7.06			
23									sy =	0.98942
24										
25									gsd =	2.69
26									X95 =	24.28
27										

1 e2: =AVERAGE(E12:E18)
2 e3: =NORMSINV(C3/(C2))
3 e4: =NORMDIST(E3,0,1,FALSE)/(1-NORMDIST(E3,0,1,TRUE))
4 e5: =SQRT((E2-LN(C5))^ 2/(E4-E3)^ 2)
5 e6: =(1-NORMDIST(E3-E5/C2,0,1,TRUE))/(1-NORMDIST(E3,0,1,TRUE))
6 g5: =C2/C3*NORMDIST(E3-E5,0,1,TRUE)*EXP(-E5*E3+(E5) ^2/2)
7 i5: =EXP((-(C2-C3)*C2)/C3*LN(E6)-E5*E3-(C2-C3)/(2*C3*C2)*(E5)^ 2)
8 e12: =LN(C12) [copy to cells e13 to e18]
9 g9: =$G$5*C9 [copy to cells g10 and g11]
10 g12: =C12 [copy to cells g13 to g18]
11 i9: =$I$5*C9 [copy to cells i10 and i11]
12 i12: =C12 [copy to cells i13 to i18]
13 j9: =LN(I9) [copy to cells j10 to j18]
14 g20: =AVERAGE(G9:G18)
15 g22: =STDEV(G9:G18)
16 j20: =AVERAGE(J9:J18)
17 j21: =EXP(J20)
j23: =SQRT(2*C2/(C2-1)*LN(G20/J21))
j25: =EXP(J23)
j26: =EXP(LN(J21)-(E5)^ 2/(2*C2)+1.645*J23)

Note: Several of the functions in the above cell formulae may be specific to Microsoft Excel. The Normsinv function calculates a fraction from 0 to 1 calculates the corresponding z-value.

The Normdist function can calculate either the probability density function or cumulative density function for a normal distribution.