Estimation of Average Concentration in the Presence of Nondetectable Values

Richard W. Hornung, and Laurence D. Reed
National Institute for Occcupational Safety and Health, Centers for Disease Control, Public Health Service,
U.S. Department of Health and Human Services, 4676 Columbia Parkway, Cincinnati, Ohio 45226

In the attempt to estimate the average concentration of a particular contaminant during some period of time, a certain proportion of the collected samples is often reported to be below the limit of detection. The statistical terminology for these results is known as censored data, i.e., nonzero values which cannot be measured but are known to be below some threshold.

Samples taken over time are assumed to follow a lognormal distribution. Given this assumption, several techniques are presented for estimation of the average concentration from data containing nondetectable values. The techniques proposed include three methods of estimation with a left-censored lognormal distribution: a maximum likelihood statistical method and two methods involving the limit of detection. Each method is evaluated using computer simulation with respect to the bias associated with estimation of the mean and standard deviation. The maximum likelihood method was shown to produce unbiased estimates of both the mean and standard deviation under a variety of conditions. However, this method is somewhat complex and involves laborious calculations and use of tables. Two simpler alternatives involve the substitution of L/2 and a new proposal of $L/\sqrt{2}$ for each nondetectable value, where L = the limit of detection. The new method was shown to provide more accurate estimation of the mean and standard deviation than the L/2 method when the data are not highly skewed. The L/2 method should be used when the data are highly skewed (geometric standard deviation [GSD] approximately 3.0 or greater). Hornung, R.W.; Reed, L.D.: Estimation of Average Concentration in the Presence of Nondetectable Values. App. Occup. Environ. Hyg. 5:46–51; 1990.

Introduction

One of the most common problems facing the industrial hygienist in characterizing data collected in a survey is a valid way of dealing with nondetectable values. Here, nondetectable is defined as any sample which is reported to be less than some value L, which is the limit of detection as defined by the sampling and analytical method. Concentrations of industrial contaminants are generally much lower today than those encountered 15 or 20 years ago. Lower average concentrations result in a higher percentage of values below the limit of detection. This problem is partially offset by improvements in analytical methods which permit lower levels to be quantified. However, despite these improvements, it has been our experience that the proportion of nondetectable samples in typical industrial hygiene data sets appears to be increasing. Therefore, a method for handling nondetectable values when they comprise a sizeable proportion of a set of samples is essential in producing accurate descriptive statistics.

The most commonly used descriptors for any data set are the mean and standard deviation. When samples are collected over time, as with grab samples within a day or personal samples on a series of days, the data are generally assumed to follow the lognormal distribution.[1,2] The corresponding parameters of this distribution are the geometric mean and geometric standard deviation. This article will be confined to the lognormal distribution. Since the logarithms of data from this distribution follow a normal distribution, the procedures described subsequently can be easily applied to data from either of these two distributions.

There are a number of techniques currently in practice for dealing with nondetectable values. Two of the more simplistic procedures are simply to ignore nondetectables or to set them equal to zero. Both result in an obvious bias when estimating the mean concentration. By ignoring or omitting nondetectables from all calculations, the estimate of the mean is biased too high. By setting them equal to zero, the estimated mean is too low. Since these practices are obviously incorrect, they will be ignored when assessing the accuracy of other procedures.

Two procedures which have been used and discussed in more recent times are a complex statistical procedure originally suggested by Hald[3] (Method 1) and a simple approximation attributed to Nehls and Akland[4] (Method 2). The Method 1 technique proposed by Hald is a maximum likelihood procedure that will produce very accurate estimates of both the mean and standard deviation of data from a censored normal distribution where the censoring point (limit of detection) is a known constant. A censored sampling distribution is one for which the only information on some of the samples is that the true measurement is

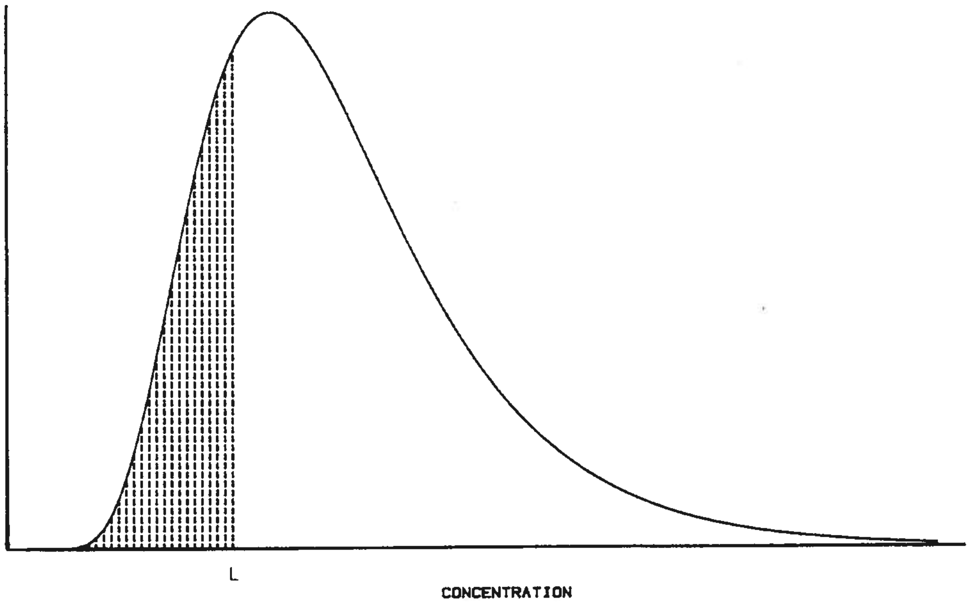

FIGURE 1. Lognormal distribution censored at the limit of detection (L).

less than (or greater than) some known value. Figure 1 illustrates the problem in terms of a censored lognormal distribution. Unfortunately, the Hald method is laborious, requiring extensive calculations and the use of two different sets of tables. In addition, it cannot be used when more than 50 percent of the data is nondetectable. Another maximum likelihood technique, attributable to Cohen,[6] can be used when more than 50 percent of the data is censored. Since this method is conceptually similar to Hald's method, it will not be compared directly to the other methods discussed in this article. The results would be equivalent since both the Hald and Cohen methods provide unbiased estimates of the mean and standard deviation.

Method 2 is an approximation recommended by Nehls and Akland that simply involves setting all nondetectables to one-half the limit of detection (L). This is reasonable since the true concentration must be somewhere between zero and L. This will subsequently be referred to as Method 2 or the "L/2 method."

The purpose of this article is to examine the accuracy of the L/2 method compared with the Hald method for varying degrees of censoring. A third method is proposed and compared with each of the other two procedures for calculating both the mean and standard deviation.

Methods

Hald Method (Method 1)

Hald's method is a mathematical technique which makes use of knowledge of the normal distribution to extrapolate back from the censoring point (the limit of detection) to produce maximum likelihood estimates of the mean and standard deviation. Hald assumes that out of *n* observations, *a* of them are below the limit of detection, L, which is constant for all *n* samples. Since a lognormal distribution is assumed, the first step is to subtract the natural log of L from the log of each of the *n–a* detectable values which results in:

$$x_i = \ln y_i - \ln L, \quad i = 1, \ldots, n-a$$

The following equations are then used to estimate the mean and standard deviation:

$$y = \frac{(n-a)\Sigma x_i^2}{2(\Sigma x_i)^2} \quad i = 1, \ldots, n-a$$

$$h = a/n = \text{proportion of nondetectable values}$$

$$z = f(h,y) \text{ which is determined from Table X in Hald's text.}$$

Then the standard deviation of the logs is estimated by

$$s = \frac{\Sigma x_i}{n-a} g(h,z)$$

$$\text{where: } g(h,z) = \frac{n-a}{a\psi(z)-(n-a)z}$$

and $\psi(z)$ is found in Hald, Table X, Part 2. The mean of the logs is then estimated by

$$\bar{x} = -zs + \ln L$$

and, as before, GM = exp ($\bar{x}$) and GSD = exp(s).

While this is a very accurate method, it is seldom used because of the laborious calculations involved. This method is presented more as a standard for comparison to the other methods than an actual recommendation.

The L/2 Method (Method 2)

This method is very simple to use and, therefore, is probably more often employed than any other procedure. All samples determined to be nondetectable are simply assigned the value of one-half the limit of detection. The log transformation is then applied to all data, and estimates of the geometric mean and geometric standard deviation are obtained as usual.

An implicit assumption of this technique is that data below the limit of detection follow a uniform distribution, i.e., every value between zero and L has an equal probability of occurring. Figure 2 illustrates the shape of such a distribution.

If industrial hygiene data actually follow such a lognormal/uniform distribution, then the L/2 method would give acceptable estimates of the mean and standard deviation. However, if the lognormal assumption is more nearly correct, this method has limitations which are addressed in the next section.

New Proposed Method (Method 3)

The L/2 method, as described in the previous section, is very simple to use but assumes a uniform distribution of samples below the limit of detection. It seems unlikely that the shape of the distribution in the left tail would depart so dramatically from the overall parent distribution of observations above the limit of detection. When the proportion of nondetectables is such that the limit of detection is not greater than the mode (the value for which the distribution has its peak), the general shape of the left side of a lognormal distribution is better approximated by a right triangle than the rectangle shown in Figure 2. Figure 3 illustrates the triangular approximation to the left of the limit of detection.

If the right triangle better approximates the lognormal distribution in this area, then the best estimate for the nondetectable values would be some value ℓ between 0.0 and L. The value of ℓ has the property that one-half the area of the triangle is to its left and one-half to its right. This can be expressed mathmetically as:

$$\int_0^{\ell} bx\,dx = 1/2\int_0^{L} bx\,dx$$

where : b = the slope of the diagonal line segment.

The integration can be solved simply as:

$$\frac{\ell^2 b}{2} = 1/2\frac{L^2 b}{2}$$

$$\ell^2 = \frac{L^2}{2}$$

$$\ell = \frac{L}{\sqrt{2}}$$

The interesting fact about this result is that if the area to the left of the limit of detection is better approximated by a triangle (compared to a rectangle), then a better estimate of the true value of a nondetectable result is $L/\sqrt{2}$ rather than L/2.

Comparative Results Using Computer Simulation

Since data for all comparisons are assumed to be lognormal and Method 1 requires a normal distribution, all calculations are performed using the natural logarithm (base e) of each observation. The estimated geometric mean and geometric standard deviation are defined as:

$$GM = \exp(\mu)$$
$$GSD = \exp(\sigma)$$

where: μ = the mean of the log transformed data
σ = the standard deviation of the log transformed data.

In order to evaluate the relative accuracy of the three methods described in the previous section, 100,000 computer-generated, random observations from each of three lognormal distributions were used.[5] The effect of the degree of censoring on accuracy was tested by setting the proportion of nondetectables at four differ-

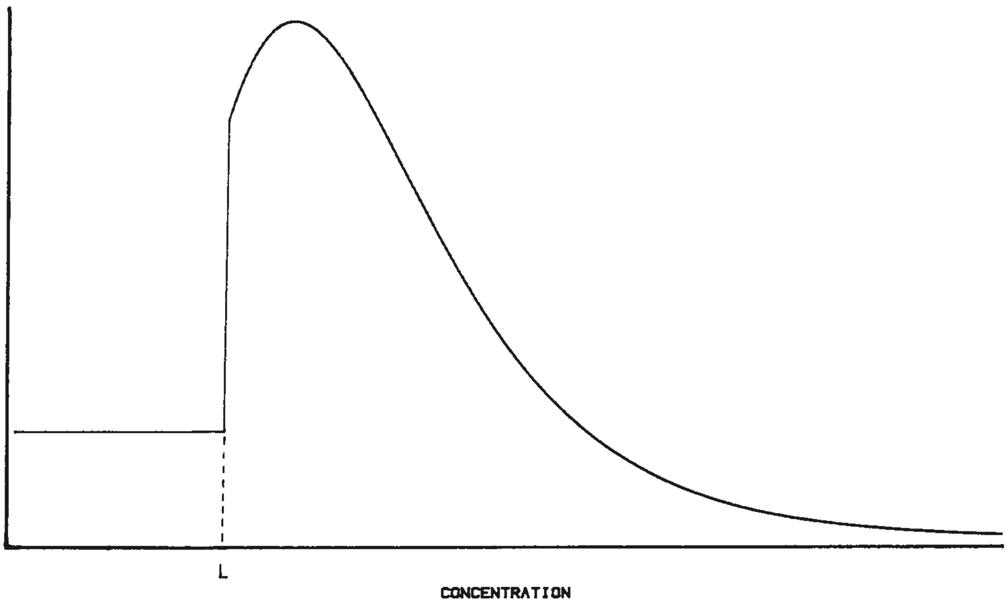

FIGURE 2. Assumed distribution for valid use of the L/2 approximation.

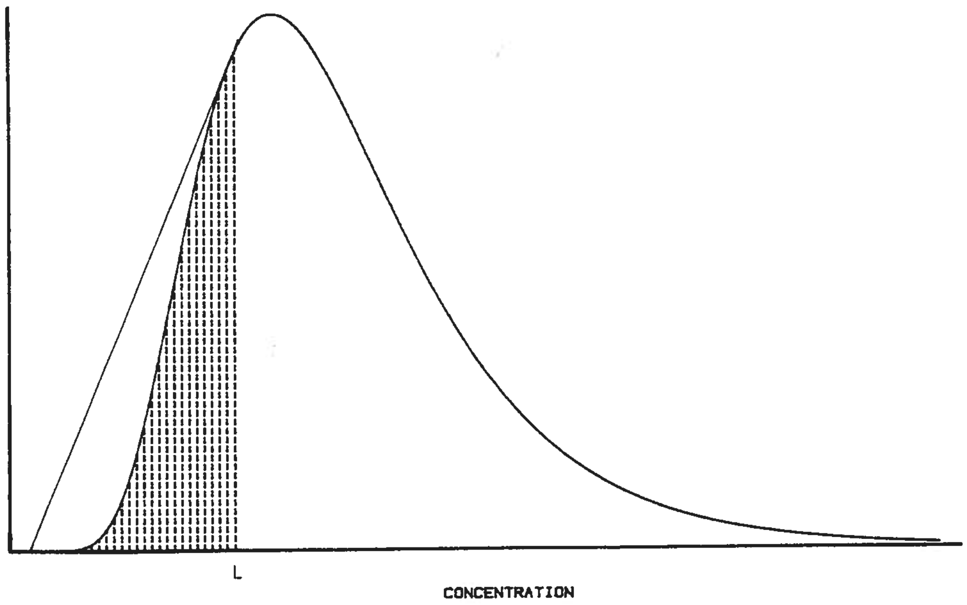

FIGURE 3. Triangular approximation used for Method 3.

ent levels: 15, 30, 45, and 60 percent. Since there was an indication that the results may have been influenced by the degree of variability in the data, each set of four censoring levels was examined for four distributions corresponding to geometric standard deviations of 1.5, 2.0, 2.5, and 3.0. These correspond to relative standard deviations of roughly 40, 80, 115, and 150 percent which is a range commonly encountered in industrial hygiene data. Since these comparisons are independent of the magnitude of the geometric mean, all three distributions were fixed with GM = 1.0.

Estimation of Geometric Mean

Estimates of the geometric mean and standard deviation were computed using each of the three methods described in the "Methods" section. Table I gives the bias associated with the estimate of the geometric mean (GM = 1.0 implies 0.0 bias) for each method. Inspection of this table reveals some interesting results. As expected, Method 1, while cumbersome to use, produces estimates of the geometric mean which are unbiased for all degrees of variability and proportions of nondetectables. It must be reiterated that this method cannot be used for more than 50 percent nondetectables. A similar method was published by Cohen[6] which provides tables for use when more than 50 percent of the data is nondetectable. However, reliability of estimates of the mean and standard deviation under these conditions is extremely poor. The only limitations to the accuracy of the Hald method are sampling error and interpolation of values in the two tables used. Methods 2 and 3, however, produce differing degrees of accuracy depending upon both the variability in the lognormal distribution (size of the GSD) and the proportion of nondetectable values.

Method 3 was clearly superior to Method 2 for low to moderate variability at all levels of censoring (percentage of nondetectables) below 45 percent. Method 2 showed a steady decline in accuracy with an increasing proportion of nondetectables at GSD = 1.5. However, at the highest level of variability (GSD = 3.0), Method 2 was superior to Method 3 at all levels of censoring.

Estimation of Geometric Standard Deviation

As an estimate of the GSD, the Hald method (Method 1) was again unbiased at all levels of variability and proportions of nondetectables. It is clearly the best method for estimating both the mean and standard deviation if accuracy is paramount. As with estimation of the geometric mean, Methods 2 and 3 produce no clear-cut favorite in estimating the geometric standard deviation. Table II details the amount of bias in these methods for all combinations of variability and proportion of nondetectables.

Method 3 produces substantially better estimates of the

TABLE I. Percent Bias in Estimating the Geometric Mean for Hald Method, Method 2, and Method 3

%	Geometric Standard Deviation			
Nondetectable	1.5	2.0	2.5	3.0
0	0.05	−0.04	−0.2	0.02
15	0.4 (1)	0.5 (1)	0.0 (1)	−0.4 (1)
	−7.2 (2)	−5.2 (2)	−3.2 (2)	−1.8 (2)
	−2.2 (3)	−0.1 (3)	1.9 (3)	3.4 (3)
30	−0.3 (1)	0.2 (1)	−0.1 (1)	0.2 (1)
	−12.3 (2)	−7.4 (2)	−3.8 (2)	11.2 (2)
	−2.6 (3)	2.8 (3)	6.9 (3)	11.2 (3)
45	0.3 (1)	−0.1 (1)	0.2 (1)	0.2 (1)
	−16.0 (2)	−7.2 (2)	0.3 (2)	6.1 (2)
	−1.8 (3)	8.4 (3)	17.1 (3)	24.1 (3)
60	—	—	—	—
	−17.9 (2)	−4.2 (2)	8.4 (2)	19.8 (2)
	1.1 (3)	17.9 (3)	33.4 (3)	47.3 (3)

TABLE II. Percent Bias in Estimating the Geometric Standard Deviation for Hald Method, Method 2, and Method 3

%	Geometric Standard Deviation			
Nondetectable	1.5	2.0	2.5	3.0
0	0.12	−0.04	−0.01	−0.05
15	0.3 (1)	0.3 (1)	−0.3 (1)	0.4 (1)
	13.4 (2)	8.1 (2)	4.2 (2)	1.0 (2)
	2.9 (3)	−1.0 (3)	−4.2 (3)	−6.6 (3)
30	−0.3 (1)	0.3 (1)	−0.6 (1)	−0.4 (1)
	16.8 (2)	7.2 (2)	0.4 (2)	−4.5 (2)
	1.7 (3)	−5.8 (3)	−11.4 (3)	−15.4 (3)
45	0.4 (1)	−0.3 (1)	0.3 (1)	0.4 (1)
	15.3 (2)	1.9 (2)	−6.9 (2)	−13.5 (2)
	−1.5 (3)	−12.2 (3)	−19.4 (3)	−24.8 (3)
60	—	—	—	—
	10.1 (2)	−6.1 (2)	−16.6 (2)	−24.2 (2)
	−6.1 (3)	−19.3 (3)	−27.9 (3)	−34.4 (3)

GSD at the lowest level of variability (GSD = 1.5) regardless of the proportion of nondetectables. At a moderate level of variability (GSD = 2.0), Method 3 is also clearly superior when the proportion of nondetectables is not greater than 30 percent. However, when the proportion is greater than 30 percent, Method 2 has less bias and then continues to surpass Method 3 for all proportions of nondetectables at the highest level of variation (GSD = 3.0).

Discussion

Comparison of the three methods for handling nondetectable values produced one clear winner. The Hald method was superior to either Method 2 (the L/2 approximation) or Method 3 (the $L/\sqrt{2}$ approximation). This comes as no surprise since this technique has previously been shown to result in unbiased estimates of the mean and standard deviation by Kushner.[7] Kushner also compared the L/2 method to Hald's method for estimation of the arithmetic mean of a lognormal distribution. Kushner found that the L/2 method agreed quite closely with Hald's method for GSD greater than 2.0 but underestimated the arithmetic mean by as much as 15 percent when the GSD was small (GSD = 1.28) and the degree of censoring was close to 50 percent. The unfortunate price that must be paid for the degree of accuracy provided by Method 1 is laborious calculation and use of specialized tables.

The real questions then are what alternatives are available and how well they work. Examination of Tables I and II suggests that Method 3 produces very accurate estimates of both the geometric mean and standard deviation when as many as half the samples are nondetectable for data of low to moderate variability. On the other hand, Method 2 produces better results for both the geometric mean and standard deviation when the data are highly skewed.

Intuitively, Method 3 should always provide better estimates than Method 2 when the data are lognormal or normal because the triangular approximation seems more realistic than a rectangular approximation. After some investigation, the reason for the superiority of Method 2 in highly skewed data was discovered. Figure 4 shows a plot of a lognormal distribution with a GSD = 3.0. Note that the mode (peak of the curve) is very close to zero. In fact, only a little over 13 percent of the data is to the left of the mode. This is evident when considering the mathematical expression for the mode:

$$
\begin{aligned}
\text{Mode} &= \exp(\mu - \sigma^2) \\
&= \text{GM}/\exp(\sigma^2)
\end{aligned}
$$

where: μ = the mean of the logs of the data
σ^2 = the variance of the logs of the data

Therefore, as σ^2 increases, the mode moves closer to zero. Since Method 3 is based upon the triangular approximation, it will only be highly accurate when the limit of detection is not much greater than the mode. Generally, this will be the case when the degree of variability is low to moderate and the proportion of nondetectables is no more than 40 percent or so.

Conclusion/Recommendation

In summary, if a high degree of accuracy is desired for both the geometric mean and geometric standard deviation, the Hald method (Method 1) should be used, but only if the proportion of nondetectables is less than 50 percent. If one is interested in producing accurate confidence limits on the geometric mean or in testing hypotheses, the Hald technique may be worth the extra effort.

In most cases, however, a simpler method will be both desirable and sufficiently accurate. When the data are not highly skewed, replacement of nondetectable values by $L/\sqrt{2}$ should produce very good estimates of both the geometric mean and standard deviation. The L/2 approximation appears warranted only when the data are highly skewed (GSD approximately 3.0 or greater). If the degree of skewness is not obvious from a quick examination of the data, the use of a histogram is suggested. Most statistical computer packages include this graphical way of examining the data (frequency of data in selected concentration intervals). Create a histogram of the detectable data only. If the frequency of the data steadily declines in every interval, Method 2 (L/2) should be used. If the frequency in the first or second interval is less than one or more of the subsequent intervals, then Method 3 ($L/\sqrt{2}$) should be used.

The use of any of these three methods when much more than half of the data is nondetectable is either not possible or will result in biased or very imprecise estimates of the geometric mean and geometric standard deviation. When the majority of samples are below the limit of detection, reporting a mean and standard deviation is a questionable practice. A better description of the data may simply be obtained by reporting the percentage of the samples below the limit of detection and the range of the remaining samples.

However, if there is a compelling reason to report a mean concentration level, Method 2 (L/2) should probably be used. Reporting the standard deviation under these conditions usually should be avoided since very little in-

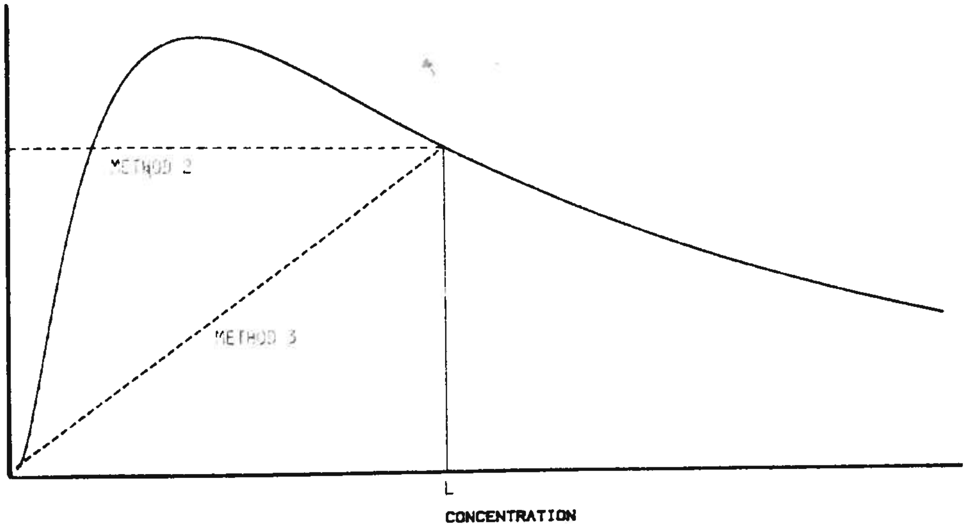

FIGURE 4. Comparison of approximations for Methods 2 and 3 when data are highly skewed. Nondetectables = 30% (GSD = 3.0).

formation is available for estimation of variability of the data.

References

1. Larsen, R.I.: A New Mathematical Model of Air Pollution Concentration Averaging Time and Frequency. J. Air Pollut. Control Assoc. 19:24 (1969).
2. National Institute for Occupational Safety and Health: Occupational Exposure Sampling Strategy Manual. DHEW (NIOSH) Pub. No. 77-173 (1977).
3. Hald, A.: Statistical Theory with Engineering Applications, pp. 144–151. John Wiley and Sons, Inc., New York (1952).
4. Nehls, G.J.; Akland, G.G.: Procedures for Handling Aerometric Data. J. Air Pollut. Control Assoc. 23:180 (1973).
5. SAS Institute, Inc.: Statistical Analysis System. Cary, NC (1985).
6. Cohen, A.C.: Tables for Maximum Likelihood Estimates: Singly Truncated or Singly Censored Samples. Technometrics 3:535 (1961).
7. Kushner, E.J.: On Determining the Statistical Parameters for Pollution Concentration from a Truncated Data Set. Atmos. Environ. 10:975 (1976).

Received 12/21/88; review decision 2/27/89; revision 7/26/89; accepted 8/10/89