Exposure estimation in the presence of nondetectable values: Another look
Finkelstein, Murray M;Verma, Dave K
AIHAJ; Mar/Apr 2001; 62, 2; ProQuest
pg. 195

AIHAJ 62:195–198 (2001)

Ms. #173

Exposure Estimation in the Presence of Nondetectable Values: Another Look

AUTHORS
Murray M. Finkelstein
Dave K. Verma

Program in Occupational Health and Environmental Medicine, McMaster University, 1200 Main St. West, Hamilton, Ontario, L8N 3Z5, Canada

A common problem faced by industrial hygienists is the selection of a valid way of dealing with those samples reported to contain nondetectable values of the contaminant. In 1990, Hornung and Reed compared a maximum likelihood estimation (MLE) statistical method and two methods involving the limit of detection, L. The MLE method was shown to produce unbiased estimates of both the mean and standard deviation under a variety of conditions. That method, however, was complicated, requiring difficult mathematical calculations. Two simpler alternatives involved the substitution of L/2 or $L/\sqrt{2}$ for each nondetectable value. The $L/\sqrt{2}$ method was recommended when the data were not highly skewed. Although the MLE method produces the best estimates of the mean and standard deviation of an industrial hygiene data set containing values below the detection limit, it was not practical to recommend this method in 1990. However, with advances in desktop computing in the past decade the MLE method is now easily implemented in commonly available spreadsheet software. This article demonstrates how this method may be implemented using spreadsheet software.

Keywords: analytical detection limits, exposure concentration, hygiene surveys, maximum likelihood estimation, missing data

A common problem faced by industrial hygienists in characterizing the data they collect in a survey is selecting a valid way of dealing with those samples reported to contain nondetectable amounts of the contaminant. Those samples are reported to have a value less than L, where L is the limit of detection as defined by the sampling and analytical methods. Concentrations of industrial contaminants are commonly much lower now than in the past, resulting in a higher percentage of values below the detection limit. In a recent survey of occupational exposure to diesel exhaust in the railroad environment, for example, Verma[1] and colleagues found that 5 of 9 samples for respirable combustible dust on board locomotives were below the limit of detection, as were 7 of 14 in the heavy repair yard.

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 over a series of days, the data generally are assumed to follow the lognormal distribution.[2] Research has shown that this model does apply to data collected in the field. An example of such a study is one that involved 82 long-term and 111 short-term personal samples for hydrocarbon exposures at petroleum bulk terminals and agencies.[3]

In 1990, Hornung and Reed[4] published an analysis of methods for estimating the descriptors of a data set in the presence of nondetectable values. The techniques proposed included a maximum likelihood estimation (MLE) statistical method and two methods involving the limit of detection. Computer simulation was used to evaluate each method 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, that method was described as "somewhat complex and requiring laborious calculations and use of tables."[4] Two simpler alternatives involved the substitution of L/2 and a new proposal for the substitution of $L/\sqrt{2}$ for each nondetectable value. The $L/\sqrt{2}$

Copyright 2001, AIHA

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

method was recommended when the data were not highly skewed. In a recent review of methods for studying the determinants of exposure, Burstyn and Teschke[5] found that hygienists reported a variety of methods for dealing with exposures below the detection limits. These included dichotomization of the exposure variable (exposed/unexposed), substitution of L/2 or L/$\sqrt{2}$ for the nondetected values, or a random selection of a value between 0 and the detection limit.

Hygienists also may be faced with a variable detection limit, as when sampling time varies. In this situation substitution of L/2 or L/$\sqrt{2}$ fails completely because these values will often be much larger than the detection limit for a sample with longer sampling time. Suppose, for example, that the detection limit is 10 units when sampling for 8 hours. If a task is sampled for only 2 hours, then the detection limit will be 40 units. It would certainly not make sense to impute a sampling result of 20 units in this circumstance. The MLE method continues to be valid in this situation. The bias in the estimation of the statistical parameters is larger in this circumstance than with a single detection limit (simulations not shown), but it is possible to compute estimates for the statistical parameters.

Although the MLE method produces the best estimates of the mean and standard deviation of an industrial hygiene data set containing values below the detection limit, it was not practical to recommend this method in 1990. Advances in desktop computing in the past decade, however, have eliminated the requirement for "laborious calculations and the use of tables." MLE is easily implemented in commonly available spreadsheet software such as Microsoft® Excel or Corel's® Quattro Pro. This article demonstrates how this method may be implemented using spreadsheet software. Once the spreadsheet template is set up, it can be readily used for any hygiene data set.

One is commonly interested in the mean of the hygiene data, μ_d, and not the mean of the logarithms of the data. Once MLE estimates of the mean and standard deviation of the logarithms of the data have been calculated, the mean of the observed data can be computed from Formula A:

$$\text{mean (observed data)} = \exp(\mu + 0.5\sigma^2) \quad \textbf{(A)}$$

The standard deviation of the observed data can be computed from Formula B:

$$\text{Standard deviation (observed data)} = ([\exp(2\mu + \sigma^2)][\exp(\sigma^2 - 1)])^{0.5} \quad \textbf{(B)}$$

Perkins[6] points out that these estimates can be biased if μ and σ are calculated from the data, as they are likely to be in most hygiene sampling situations. He presents a tabular method to compute a minimum variance unbiased estimator. Calculation of the confidence limits on the mean is rather complicated. For details see the discussion in Perkins.[6, p. 334]

AN OVERVIEW OF MLE

MLE is a method for estimating the parameters of a statistical distribution from observed data. The parameters selected are those that would maximize the probability of observing the data if they were randomly drawn from the statistical distribution. To simplify the discussion, only the case of the lognormal distribution will be considered.

The lognormal is a distribution with two parameters, the mean, μ, which specifies the center of the distribution; and the standard deviation, σ, which specifies the spread of the data. Let $\ln(x_i)$ be the logarithm of the measured value, x, of hygiene sample i. Then, the probability distribution is defined by:

$$\text{Lognormal probability function } f(x_i, \mu, \sigma) = \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(\ln(x_i) - \mu)^2}{2\sigma^2}\right]$$

which is the probability of observing a particular value x_i, given μ and σ for the distribution. Now, one of the properties of probability is that, if P_1, P_2, and P_3 are the probabilities of observing individual events, such as "heads" in the flip of a coin, then the probability of observing the three events together is the product of the individual probabilities $P_1 \times P_2 \times P_3$ (so that the probability of observing three heads in the toss of a coin is $0.5 \times 0.5 \times 0.5 = 0.125$). If there are thus three observations, $y_1 = \ln(x_1)$, $y_2 = \ln(x_2)$, and $y_3 = \ln(x_3)$, from a lognormal distribution with mean, μ, and standard deviation, σ, the probability of obtaining these values for the three observations is

$$P(x_1, x_2, x_3 \mid \mu, \sigma) = \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(y_1 - \mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(y_2 - \mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(y_3 - \mu)^2}{2\sigma^2}\right]$$

In statistical terminology this probability is called a *likelihood*, and the method of maximum likelihood finds those values of μ and σ that maximize this probability. If the only information available about the distribution of exposures in a workplace comes from these observations, the mean and standard deviation of the exposures can be estimated by finding those values of μ and σ that will maximize the probability of observing these data values under the assumption that the data are lognormally distributed. The MLE method requires the specification of a probability distribution for the data. The method for the lognormal distribution has been outlined. If the data are not lognormally distributed, the mean and standard deviation will not be correct. Data analysts should thus examine the distribution of their data, using a method such as cumulative distribution plots, to confirm that the assumption of lognormality is reasonable.

Figure 1 illustrates these ideas with three data points drawn from a lognormal distribution. The vertical lines show the probability of observing each value, given a variety of means with fixed standard deviation. The text in the panels gives the likelihood for each experimental value of the mean. The likelihood is maximal in Panel C, which provides the best estimate of the mean, given the data, x.

Now, what if all we know about observation x_3 is that it is less than some detection limit, L? The probability of observing a value less than L is P_L, the area under the lognormal distribution curve up to log(L) (from the laws of probability the total area under the curve is equal to 1). This probability is available in statistical tables and is programmed into spreadsheet applications. With this value for P_L the probability of observing x_1, x_2, x_3 for the three samples, given μ and σ, is

$$\frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(y_1 - \mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[\frac{-(y_2 - \mu)^2}{2\sigma^2}\right] \times P_L(\mu, \sigma)$$

It is conventional to simplify computations by taking logarithms to convert this product to sums. The task is to find values of μ and σ that will maximize the probability (log likelihood). It turns out that this is easy to do using spreadsheet software on a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

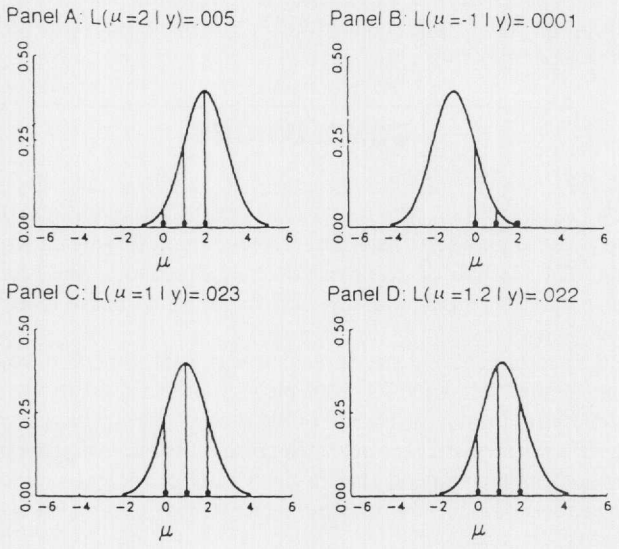

FIGURE 1. A schematic illustration of MLE. Three points have been selected from a lognormal distribution. A normal curve with variable mean is slid along the x axis. The likelihood for each observation is given by the line segment to the normal curve. The likelihood for the sample of three observations, given the selected value for the mean, is given by the sum of the individual likelihoods and is displayed in the title bar for each panel. Maximizing the sample likelihood selects the optimal estimate of the mean of the distribution.

personal computer. The next section demonstrates how that is done.

USING SPREADSHEET SOFTWARE TO COMPUTE MLES

The goal is to maximize the likelihood of the observations that are assumed to follow a lognormal distribution. It has been shown that, if all we know about observation x is that it is less than some detection limit, L, then the probability of observing a value less than L is P_L, the area under the normal distribution curve up to log(L). This probability is programmed into spreadsheet applications. In Microsoft Excel, the function is NORMDIST. This function returns the normal cumulative distribution for the specified mean and standard deviation. The syntax is: NORMDIST(x, mean, standard_dev, cumulative), where x is the value for which the probability is needed (in this instance, since a normal distribution of the logarithms of the observed values are being discussed, x is the logarithm of the measured value); mean is the mean of the lognormal distribution; standard_dev is the standard deviation of the lognormal distribution; and cumulative is a logical value that determines the form of the function. If cumulative is TRUE, NORMDIST returns the cumulative distribution function, which is the desired result.

In maximizing the likelihood function, it is conventional to apply a logarithmic transformation to convert the product of likelihood terms to a sum. Figure 2 illustrates a spreadsheet template for the maximization of the log-likelihood. For the purposes of demonstration, 10 data points have been drawn from a lognormal distribution and are listed in Column A. An analytical detection limit of three units has been presumed, and so the smallest data

	A	B	C	D	E
1	Demonstration Spreadsheet for Maximum Likelihood Calculations				
2					
3					
4					Solver Cells
5	DATA*	Log Likelihood of Observation, given estimate of Mean & SD	Starter	Mean	2.14
6	20.25	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A6)-E$5)/E$6)^2))	Starter	SD	0.71
7	19.94	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A7)-E$5)/E$6)^2))			
8	9.52	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A8)-E$5)/E$6)^2))			
9	8.29	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A9)-E$5)/E$6)^2))			
10	7.23	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A10)-E$5)/E$6)^2))			
11	4.41	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A11)-E$5)/E$6)^2))			
12	3.06	=LN((1/((2*PI())^0.5*E$6))*EXP(-(1/2)*((LN(A12)-E$5)/E$6)^2))			
13	<3	=LN(NORMDIST(LN(3),E$5,E$6,TRUE))			
14	<3	=LN(NORMDIST(LN(3),E$5,E$6,TRUE))			
15	<3	=LN(NORMDIST(LN(3),E$5,E$6,TRUE))			
16					
17					
18		Total LogLikelihood			
19		=SUM(B6:B15)			
20					
21					
22	* Demonstration data are drawn from a lognormal distribution with Mean (of logs) = 2 and Geometric Standard Deviation = 3.0				
23					
24	Mean of Logarithms	of Observed data = 2.14 (Computed from observed data)			
25	Standard Deviation	of Logarithms of Observed data = 0.71 (Corresponds to GSD of 2.03)			
26					
27	Estimated Mean of	Observed data =EXP(E5 + 0.5*E6^2)			
28	Estimated Standard	Deviation of Observed Data=(EXP(2*E5+E6^2)*((EXP(E6^2) - 1)))^0.5			

FIGURE 2. A sample spreadsheet demonstrating the structure of a template for MLE

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

values have been labeled as <3. To illustrate the underlying formula structure of the spreadsheet, turn on the formula view in Excel.

For the data points with measured values, Column B contains the logarithm of the likelihood for that observation. A crucial point is that one cannot work with symbolic equations in spreadsheet software. Numerical values must thus be substituted for μ and σ. Thus, an initial, or starter, value is inserted for the mean and standard deviation into the spreadsheet. These appear in cells E5 and E6. The Solver Module in the spreadsheet software will vary the values in E5 and E6 to maximize the sum of the log likelihoods in Column B. Convenient starting values for the estimates in E5 and E6 are the mean and standard deviation of the logarithms of the observed data values in Column A.

Now, all that is known about the smallest data values is that they are less than 3. The likelihoods for the logarithms of these 3 data points are given by the area under the normal distribution curve up to log(3). These are computed by the NORMDIST spreadsheet function. The value of x is ln(3), the mean and standard deviation are given by the starter values in E5 and E6, and "cumulative" is set to TRUE, since we want the area under the curve up to ln(3), given μ and σ. The goal is to select values for E5 and E6 to maximize the sum of the likelihoods, which can be found in cell B19.

To maximize the sum of the likelihoods, select the Solver Tool from the Tools menu and then select a target cell. Here, it is B19, the sum of the log likelihoods. Select the option to maximize the value of B19. Then select the cells to change in order to achieve this maximization. Here, they are E5 and E6, the initial estimates of μ and σ. Clicking on the Solve button initiates the computation, and the values in E5 and E6 are replaced with 1.64 and 0.97, the maximum likelihood estimates of the mean and standard deviation of the logarithms of data from the parent distribution of the data points in Column A. Rows 27 and 28 contain the formulae (A and B above) for the computation of the mean, μ_d, and standard deviation, σ_d, of the hygiene data from the lognormal mean and standard deviations.

By substituting hygiene data values for the example data in this template, MLE estimates can be computed for any data set. If the measurements involve more than one detection limit, then the appropriate limits may be substituted for ln(x) in the NORMDIST spreadsheet function.

CONCLUSIONS

When hygiene samples are collected over time, as with grab samples within a day or personal samples over a series of days, the data generally are assumed to follow the lognormal distribution. The method of maximum likelihood produces the best estimates of the mean and standard deviation in many industrial hygiene datasets.

The authors have demonstrated how the MLE method may be implemented in commonly available spreadsheet software (an Excel template for the calculation of MLE estimates is posted at http://www.fhs.mcmaster.ca/oehl). Because of its optimal properties, the authors recommend that hygienists adopt the MLE method when their data include measurements reported to be below the limit of detection.

REFERENCES

1. **Verma, D.K., L. Shaw, J.A. Julian, K. Smolynec, C. Wood, and D.A. Shaw:** A comparison of sampling and analytical methods for assessing occupational exposure to diesel exhaust in a railroad work environment. *Appl. Occup. Environ. Hyg. 14*:701–714 (1999).
2. **Leidel, N.A., K A. Busch, and J.R. Lynch:** *Occupational Exposure Sampling Strategy Manual* (DHEW [NIOSH] Publication no. 77–173). Cincinnati, OH: National Institute for Occupational Safety and Health, 1977.
3. **Verma, D.K., J. A. Julian, G. Bebee, W. K. Cheng, K. Holborn, and L. Shaw:** Hydrocarbon exposures at petroleum bulk terminals and agencies. *Am. Ind. Hyg. Assoc. J. 53*:645–656 (1992).
4. **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).
5. **Burstyn, I., and K. Teschke:** Studying the Determinants of exposure: A review of methods. *Am. Ind. Hyg. Assoc. J.* 60:57–72 (1999).
6. **Perkins, J.L.:** Quantitative surveying—application of distributional models to exposure assessment. In *Modern Industrial Hygiene*, vol. 1. New York: Van Nostrand Reinhold, 1997. pp. 318–353.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.