

AUTHORS

Daniel J. Strom
Paul S. Stansbury

Risk Analysis and Health
Protection Group, Pacific
Northwest National Laboratory,
Battelle Boulevard, P.O. Box
999, Richland, Washington
99352; e-mail:
daniel.j.strom@pnl.gov

Determining Parameters of Lognormal Distributions from Minimal Information

The lognormal distribution has a number of properties that do not lend themselves to simple "back-of-the-envelope" calculations. Mathematical relationships are presented for the basic parameters of the large population lognormal distribution as a function of characteristics available to, or needed by, the risk analyst. A freeware computer program called LOGNORM4 has been written to take the tedium out of determining various characteristics of lognormal distributions, given 1 of 15 sets of values that uniquely specify a lognormal distribution.

Keywords: aerosol size distribution, computer program, environmental health, lognormal distribution, parameters, probabilistic risk assessment

In considering the health effects of a hazardous or potentially hazardous agent in the indoor or outdoor environment, the person assessing the risks often wishes to compute more than just the central estimate or just the upper estimate of the risk. To compute a distribution of health risks, the distribution of the amount of the agent to which an individual or a group of individuals is exposed must be known. Often the distribution of pollutants in the environment is lognormally distributed because many environmental processes are governed by the product of independent random variables.⁽¹⁾ However, the risk analyst may not have ready access to the entire set of original data. Instead, he or she might be able to find minimal information published, or otherwise documented, giving the mean and standard deviation (SD) with no distribution specified, or perhaps the geometric mean and geometric standard deviation (GSD) without a complete justification of why a lognormal distribution was assumed.

Often, a risk analyst must address a problem with only such minimal data. Whenever one assumes a type of underlying statistical distribution without the data to show that such an assumption is correct or at least plausible, there is a danger of drawing erroneous conclusions. Those who read this article or use the software it describes should be aware of these dangers and include appropriate caveats or warnings with the results of such analyses.

However, once a risk analyst has decided to assume a lognormal distribution and to base it

on the minimal information available, there is often another problem to face. The minimal information he or she has does not fit the analysis tools. For example, one popular probabilistic decision rule software package, Crystal Ball (v. 2.0.3, Decisioneering, Inc., Denver, Col.), requires that the user specify lognormal distributions by the mean and arithmetic (not geometric) SD, parameters that are often not published when the median and GSD are specified.

The authors of this article wish to give a second significant warning. The algebraic relationships used in this article give an accurate characterization of a lognormal distribution when the parameter estimates (that is, the statistics computed from the observed data) are derived from a large sample that is taken from an underlying distribution that is truly a single lognormal distribution. For example, the article presents methodology for computing the geometric mean and GSD of a lognormal distribution given the median (50th percentile) and the 95th percentile of the observed values. If the actual parent distribution from which the observations were made was not lognormal or if the sample size was so small that median and 95th percentile values are not accurate estimates of the true values, then the observed lognormal distribution characterized with the methods given may be highly inaccurate and the conclusions based on the distribution developed gravely wrong.

The lognormal distribution is skewed, with distinct values for the mean, median, and mode.

Pacific Northwest
National Laboratory is
operated for the U.S.
Department of Energy by
Battelle Memorial
Institute under contract
DE-AC06-76RLO 1830.

TABLE I. Symbols, Meanings, and Algebraic Relationships for the Lognormal Distribution

Symbol	Meaning	Algebraic Relationship
μ	mean (= median) of natural logarithms of x	$\ln(\bar{x})$
σ	standard deviation of logarithms of x	$\ln(\text{GSD})$
\bar{x}	median	e^μ
GSD	geometric standard deviation	e^σ
\hat{x}	mean	$e^{\mu + \sigma^2/2}$
\hat{x}	mode	$e^{\mu - \sigma^2}$
CV	coefficient of variation	$CV^2 = e^{\sigma^2} - 1$
Var	variance	$\bar{x}^2 \cdot CV^2$
SD	[arithmetic] standard deviation	$\bar{x} \cdot CV$
x_1, x_2	values 1 and 2	
Z_1, Z_2	standard normal deviates for x_1 and x_2	
p_1, p_2	percentiles for x_1 and x_2	
f_1, f_2	fractiles for x_1 and x_2	

Note: Adapted from Reference 2.

Although many of its properties can be calculated analytically, these values are neither intuitive nor straightforward for most people. To solve these problems in a convenient form, the properties of the lognormal distribution are examined and algebraic solutions developed for some common problems. The solutions are encoded in a computer program called LOGNORM4.

This program is copyrighted freeware and may be downloaded from <http://qecc.pnl.gov/LOGNORM4.htm> over the Internet.

THEORETICAL BACKGROUND

The lognormal probability density function is given by

$$P(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right], \quad (1)$$

for positive x, where μ is the mean of the logarithms of x, and σ^2 is the variance of the logarithms of x. The authors have adopted the convention used by Atchinson and Brown,⁽²⁾ in which μ and σ are characteristics of the underlying population distribution and are Greek letters. Sample functions and estimators are roman. Some of the formulas listed here are also available in other common industrial hygiene references.^(3,4)

The basic lognormal terms are defined and a number of the basic lognormal relationships are given in Table I. The first relationship in Table I is based on the property of the normal distribution that the mean is equal to the median. Thus, the mean of the natural logarithms of x corresponds to the median of the natural logarithms. The medians of the log-transformed and untransformed distributions are related by the natural log function. From an examination of Table I, fifteen conceptually distinct ways of uniquely specifying a lognormal distribution were identified, and these are listed in Table II.

For data that are known to be lognormally distributed, one may simply find the average and SD of the logarithms of the data and use relationships in Table I to determine other parameters of interest. For data suspected to be lognormally distributed, one of many fitting routines can be used, such as LOGNORM4.⁽⁵⁾ The latter program also fits lognormal distributions to censored and categorized data. The problems of dealing with censored or "less-than" data that are believed to be lognormally distributed are described elsewhere.⁽⁶⁾

TABLE II. Fifteen Conceptually Distinct Ways of Uniquely Specifying a Lognormal Distribution

Parameters Specifying a Unique Lognormal Distribution	Formulas for μ and σ
Mean and median (or their natural logs)	Table I and Equation 3
Mean and mode (or their natural logs)	Equations 4 and 14
Median and mode (or their natural logs)	Table I and Equation 5
Median (or its natural log) and the GSD	Table I
or σ	
Mean (or its natural log) and the GSD or σ	Equation 13
Mode (or its natural log) and the GSD or σ	Equation 14
A value with its percentile or fractile or standard normal deviate and the GSD or σ	Equation 16
Median and a value with its percentile or fractile or standard normal deviate	Equation 7
Mean and a value with its percentile or fractile or standard normal deviate	Equations 19 and 13
The mode and a value with its percentile or fractile or standard normal deviate	Equations 20 and 14
Median and [arithmetic] standard deviation or coefficient of variation	SD: Eqn. 9 CV: Eqn. 6
Mean and the [arithmetic] standard deviation or coefficient of variation	Equations 6 and 13
Mode and the [arithmetic] standard deviation or coefficient of variation	SD: Eqns. 10 and 11 CV: Eqns. 6 and 21
A value with its percentile or fractile or standard normal deviate and the [arithmetic] standard deviation or coefficient of variation	SD: Eqns. 12 and 16 CV: Eqns. 6 and 16
A pair of values and their percentiles or fractiles or standard normal deviates	Equations 8 and 15

When only summary measures of a data set are available, using a fitting routine is not possible. Starting from summary measures and the basic relationships in Table I, a number of formulas are derived to calculate the various characteristics of a lognormal distribution that may be of use in risk analysis.

The SD of the logarithms of x, σ , can be found directly in seven distinct ways:

$$\sigma = \ln(\text{GSD}) \quad (2)$$

$$\sigma = \sqrt{2 \ln(\bar{x}/\hat{x})} \quad (3)$$

$$\sigma = \sqrt{2 \ln(\bar{x}/\hat{x})/3} \quad (4)$$

$$\sigma = \sqrt{\ln(\bar{x}/\hat{x})} \quad (5)$$

$$\sigma = \sqrt{\ln(\text{CV}^2 + 1)} \quad (6)$$

$$\sigma = \frac{\ln(x_1/\bar{x})}{z_1} \quad (7)$$

$$\sigma = \frac{\ln(x_1/x_2)}{z_1 - z_2} \quad (8)$$

Note that $\sigma > 0$, and $\text{GSD} > 1$. There are three additional ways of finding σ that are less straightforward. The equation relating σ to \bar{x} and SD is transcendental and must be solved numerically:

$$(e^{\sigma^2/2})\sqrt{e^{\sigma^2} - 1} = \frac{\text{SD}}{\bar{x}} \quad (9)$$

The equation relating \bar{x} to \hat{x} and SD must also be solved numerically:

$$\hat{x}^2 SD^2 - \bar{x}^4 + \bar{x}^3 \hat{x} = 0. \quad (10)$$

Once this is done, σ can be found from

$$\sigma = \sqrt{\mu - \ln(\hat{x})}. \quad (11)$$

The relationship between σ , x_1 , z_1 , and the SD is

$$\frac{SD^2}{x_1^2} = \exp(-2z_1\sigma + \sigma^2)(e^{\sigma^2} - 1). \quad (12)$$

Equation 12 must be solved numerically, and yields multiple values for σ when $\sigma < 2z_1$.

In addition to the formula in Table I, the median can be found from the mean or the mode:

$$\bar{x} = \bar{x}e^{-\sigma^2/2} \quad (13)$$

$$\bar{x} = \hat{x}e^{\sigma^2}. \quad (14)$$

The value of μ can be found from two values and their standard normal deviates:

$$\mu = \ln(x_1) + \ln(x_2/x_1) \cdot \frac{(0 - z_1)}{(z_2 - z_1)}, \quad (15)$$

where the zero is the standard normal deviate of the median (z_0). The value of μ also can be found from a single value, its standard normal deviate, and σ :

$$\mu = \ln(x_1) - \sigma z_1. \quad (16)$$

The mean can be found directly from \bar{x} and σ :

$$\bar{x} = \bar{x}e^{\sigma^2/2}. \quad (17)$$

The mode also can be found directly from \bar{x} and σ :

$$\hat{x} = \bar{x}e^{-\sigma^2}. \quad (18)$$

When the mean and a value with its fractile, percentile, or standard normal deviate are known, σ can be found from

$$\sigma = \begin{cases} z_1 + \sqrt{z_1^2 + 2 \ln(\bar{x}/x_1)} & \text{for } \bar{x} < x_1 \text{ or} \\ z_1 - \sqrt{z_1^2 + 2 \ln(\bar{x}/x_1)} & \text{for } \bar{x} > x_1. \end{cases} \quad (19)$$

When the mode and a value with its fractile, percentile, or standard normal deviate are known, σ can be found from

$$\sigma = \begin{cases} \frac{-z_1 + \sqrt{z_1^2 - 4 \ln(\hat{x}/x_1)}}{2} & \text{for } \hat{x} > x_1 \text{ or} \\ \frac{-z_1 - \sqrt{z_1^2 - 4 \ln(\hat{x}/x_1)}}{2} & \text{for } \hat{x} < x_1. \end{cases} \quad (20)$$

When the mode and CV are given, μ is

$$\mu = \ln(\hat{x}) + \ln(CV^2 + 1). \quad (21)$$

The variance is given by

$$\text{Var} = SD^2, \quad (22)$$

or

$$\text{Var} = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1). \quad (23)$$

The skewness is

$$\text{skewness} = CV^3 + 3 CV. \quad (24)$$

The kurtosis is

$$\text{kurtosis} = CV^8 + 6CV^6 + 15 CV^4 + 16 CV^2. \quad (25)$$

The standard normal deviate of the mean is

$$z_x = \sigma/2, \quad (26)$$

whereas that of the mode is

$$z_x = -\sigma. \quad (27)$$

The standard normal deviate of a value x can be found from

$$z_x = \frac{\ln(x) - \mu}{\sigma}, \quad (28)$$

whereas the inverse operation of finding an x value from a standard normal deviate is

$$x = e^{\mu + z_x \sigma}. \quad (29)$$

The transformation from fractiles or percentiles to standard normal deviates and vice versa is done using tables of the cumulative standard normal distribution.

The relationships given above are useful for data sets for which uniquely defining characteristics are known. If there is more than one set of summary statistics sufficient to calculate estimates of μ and σ , they should be calculated each way possible and the results should be in good agreement in order to conclude that the distribution is indeed lognormal.

SOFTWARE LOGNORM4 AND EXAMPLES

These equations were all coded into a program called LOGNORM4 written in QuickBasic 4.5 (Microsoft Corp., Redmond, Wash.) or QBasic, which comes with MS-DOS 5.0 (Microsoft Corp.). LOGNORM4 works fine in a DOS window under Windows[®] 95 or Windows 98 (Microsoft Corp.). Using the cumulative density function of the standard normal distribution, the program determines all unspecified parameters for a lognormal distribution given any unique set listed in Table II. Once the parameters μ and σ are determined, LOGNORM4 finds any other percentiles, fractiles, standard normal deviates, or values, given one of these four.

The example of natural uranium concentrations in urine for a group of unexposed persons has occurred in two articles recently. Beyer and colleagues⁽⁷⁾ give $\bar{x} = 11.4$ ng/d, $\bar{x} = 8.7$ ng/d, and $SD = 9.7$ ng/d. From these data, an assumed lognormal distribution can be constructed in three ways (Table II, rows 1, 11, and 12). Each results in a GSD of 2.09. Créhange and Gerasimo⁽⁸⁾ give a mean of 0.57 $\mu\text{g/L}$ with an SD of 0.27 $\mu\text{g/L}$. From these data, an assumed lognormal distribution with $\bar{x} = 0.52$ $\mu\text{g/L}$ and $GSD = 1.57$ is characterized using the Table II, row 12 method.

LOGNORM4 contains many basic properties of the lognormal distribution in addition to those in the equations above, such as

$$\bar{x} > \bar{x} > \hat{x} \quad (30)$$

that prevent the entry of combinations of parameters inconsistent with a lognormal distribution. For example, the choices $\bar{x} = 60$ and $x_{95\%ile} = 300$ result in the radicand in Equation (19) being negative. These choices are thus inconsistent with a lognormal distribution.

SUMMARY AND CONCLUSIONS

Risk analysts make frequent use of the lognormal distribution. Mathematical relationships are presented for the basic parameters of the large-population lognormal distribution as a function of characteristics available to or needed by the risk analyst. The LOGNORM4 computer program was written to take the tedium

out of determining various characteristics of lognormal distributions from minimal information. Fifteen different ways of uniquely specifying a lognormal distribution are presented.

ACKNOWLEDGMENTS

The authors wish to thank James W. Hardin for critical review and discussion.

REFERENCES

1. Ott, W.R.: *Environmental Statistics and Data Analysis*. New York: Lewis Publishers, 1995.
2. Aitchison, J., and J.A.C. Brown: *The Lognormal Distribution*. Cambridge, U.K.: Cambridge University Press, 1957.
3. Leidel, N.A., K.A. Busch, and J.R. Lynch: *Occupational Exposure Sampling Strategy Manual* (NIOSH publication no. 77-173; NTIS publication no. PB274792). Springfield, Va.: National Technical Information Service, 1977.
4. Rappaport, S.M.: Interpreting levels of exposures to chemical agents. In *Patty's Industrial Hygiene and Toxicology*, vol. 3, part A, 3rd ed., R.L. Harris, L.J. Cralley, and L.V. Cralley (eds.). New York: John Wiley & Sons, 1994.
5. Strom, D.J.: LOGNORMML. *RSIC Newsletter* 325:5 (December 1991). (Radiation Shielding Information Center Code No. PSR-3 70)
6. Strom, D.J.: Estimating individual and collective doses to groups with "less than detectable" doses: A method for use in epidemiologic studies. *Health Phys.* 51:437-445 (1986).
7. Beyer, D., R. Giehl, and G. Pilwat: Normal concentration of uranium in urine. *Health Phys.* 64:321 (1993).
8. Créhange, G., and P. Gerasimo: Méthode rapide de dosage de l'uranium dans les urines. *Radioprotection* 27:283-290 (1992).