Mathematical Properties of the Risk Equation When Variability is Present

David E. Burmaster
Alceon Corporation
PO Box 382669 Harvard Square Station
Cambridge, MA 02238-2669
tel: 617-864-4300; fax: 617-864-9954
email: deb@Alceon.com

Leslie R. Bloomfield
Stone & Webster Engineering Corporation
245 Summer Street
Boston, MA 02210
tel: 617-589-1765; fax: 617-589-2922
email: leslie.bloomfield@swec.com

Abstract

When random variables are used to represent variability, the risk equation has mathematical properties poorly understood by many risk assessors. Variability represents the heterogeneity in a well-characterized population, usually not reducible through further measurement or study. We follow the lead of most mathematicians in using random variables to represent and analyze variability. To illustrate the issues, we use LogNormal distributions to model variability.

1.0 Introduction

When estimating the incremental lifetime cancer risk, R, from an environmental exposure to a single carcinogenic chemical via a single exposure pathway, risk assessors often use equations of this fundamental form:

$$R = \frac{\prod_{i=1}^{I} X_i}{\prod_{j=1}^{J} Y_j} \qquad \text{Eqn 1}$$

where $\prod$ indicates a product over the index. In common practice, risk assessors use point values (i.e., real numbers) for each variable in Eqn 1. Burmaster and Thompson (1995a, b) have discussed the origins and interpretation of Eqn 1 in deterministic risk assessments.

Most risk assessors now agree that all the variables in Eqn 1 contain both (i) variability and/or (ii) uncertainty. In this discussion, variability represents the heterogeneity in a well-characterized population [and is usually not reducible through further measurement or study] while uncertainty represents our

ignorance about a poorly-characterized phenomenon or models [and may be reducible through further measurement or study]. Thus, variability is a property of the natural system under analyst, while uncertainty is a property of the analyst. Here, we focus exclusively on variability -- not because uncertainty is unimportant, but because the introduction of variability alone illustrates the main mathematical points of this discussion.

In the probabilistic paradigm, Eqn 1 remains the fundamental equation of risk assessment (Burmaster & Thompson, 1995a, b). However, in the fully probabilistic framework, each of the variables in Eqn 1 is a positive random variable represented by a probability density function (PDF) or a cumulative distribution function (CDF) (see, e.g., Feller, 1968 & 1971). To emphasize this change in perspective, we re-write Eqn 1 as Eqn 2, with doubly underscored symbols to denote that each variable is now a random variable that expresses variability in a quantity. We also create Eqns 3 and 4, each an alternative and equivalent representation of Eqn 2:

$$\underline{\underline{R}} = \frac{\prod_{i=1}^{I} \underline{\underline{X}}_i}{\prod_{j=1}^{J} \underline{\underline{Y}}_j} \qquad \text{Eqn 2}$$

$$\underline{\underline{R}} = f(\underline{\underline{X}}_i, \underline{\underline{Y}}_j) \quad \text{for } i = 1, ..., I \text{ and } j = 1, ..., J \qquad \text{Eqn 3}$$

$$\underline{\underline{R}} = \frac{g(\underline{\underline{X}}_i)}{h(\underline{\underline{Y}}_j)} \quad \text{for } i = 1, ..., I \text{ and } j = 1, ..., J \qquad \text{Eqn 4}$$

In Eqn 4, we use the notation $g(\underline{\underline{X}}_i)$ for the product of random variables in the numerator and the notation $h(\underline{\underline{Y}}_j)$ for the product of random variables in the denominator so we can refer to the numerator and denominator separately as needed. We will continue to denote real variables (point values) without the double underscores. With knowledge of the distributions of all the $\underline{\underline{X}}_i$ and $\underline{\underline{Y}}_j$, an analyst can calculate a closed form expression for the distribution $\underline{\underline{R}}$ in a handful of special cases with independent variables (Springer, 1979). In most practical cases, including those cases with correlated or jointly distributed random variables on the right hand side of the risk equation, the analyst can simulate a numerical approximation to the distribution $\underline{\underline{R}}$ (Rubenstein, 1981; Morgan, 1984).

2.0 Background on Two-Parameter LogNormal Distributions

LogNormal distributions with two constant parameters play a central role in expressing variability in human and ecological risk assessment for at least three reasons. First, many physical, chemical, biological, and statistical processes tend to create random variables that follow two-parameter LogNormal distributions for expressing variability (Hattis & Burmaster, 1994). For example, the physical mixing and dilution of one material (say, a miscible or soluble contaminant) into another material (say, surface water in a bay) tends to create non equilibrium concentrations which are LogNormal in character (Ott, 1990; Ott, 1995). Second, when the conditions of the Central Limit Theorem hold, the mathematical process of multiplying a series of random variables will produce a new random variable (the product) which, in the limit, is LogNormal in character, regardless of the distributions from which the input variables arise (Benjamin & Cornell, 1970). Finally, two-parameter LogNormal distributions are self-replicating under multiplication and division, i.e., products and quotients of such LogNormal random variables are themselves distributed lognormally (Aitchison & Brown, 1957; Crow & Shimizu, 1988). All these points apply to Eqns 2, 3, and 4.

The two-parameter LogNormal distribution expressing variability takes its name from the fundamental property that the logarithm of the random variable is distributed according to a Normal or Gaussian distribution (Evans et al, 1993):

$$\ln[\underline{\underline{X}}] \sim N(\mu, \sigma) \qquad \text{Eqn 5}$$

where ln[•] denotes the natural or Napierian logarithm function (base e) and N(•, •) denotes a Normal or Gaussian distribution with two constant parameters, the mean μ and the standard deviation σ (with $\sigma > 0$). In Eqn 5, $\underline{\underline{X}}$ is a two-parameter LogNormal random variable, and $\ln[\underline{\underline{X}}]$ is a Normal random variable. In Eqn 5, μ is the mean and σ is the standard deviation of the distribution for the Normal random variable $\ln[\underline{\underline{X}}]$, not the LogNormal random variable $\underline{\underline{X}}$. Many people say that Eqn 1 represents the LogNormal random variable $\underline{\underline{X}}$ "in logarithmic space." As can be seen in Eqn 5, the random variable $\ln[\underline{\underline{X}}]$ is distributed normally, but the random variable $\underline{\underline{X}}$ is distributed lognormally.

The information coded in Eqn 5 is identical to the information coded in Eqn 6:

$$\underline{\underline{X}} \quad \sim \quad \exp[\, N(\mu, \sigma) \,] \qquad \text{Eqn 6}$$

where exp[•] denotes the exponential function and N(•, •) again denotes the same Normal or Gaussian distribution with the same two constant parameters, mean μ and standard deviation σ (with $\sigma > 0$) as above. In Eqn 6, $\underline{\underline{X}}$ is a two-parameter LogNormal random variable. As earlier, μ is the mean and σ is the standard deviation of the Normal random variable $\ln[\underline{\underline{X}}]$, not the LogNormal random variable $\underline{\underline{X}}$. Many people say that Eqn 6 represents the LogNormal random variable $\underline{\underline{X}}$ "in linear space." When working with Eqn 6 as the representation for a LogNormal random variable $\underline{\underline{X}}$, many people refer to $N(\mu, \sigma)$ as the "underlying Normal distribution" or "the Normal distribution in logarithmic space" as a way to remember its origins.

3.0 The Fundamental Risk Equation With All LogNormal Random Variables

3.1 The General Case

If all the inputs to the fundamental risk equation, Eqn 2, are independent LogNormal random variables of the form:

$$\underline{\underline{X}}_i \quad \sim \quad \exp[\, N(\mu_i, \sigma_i) \,] \qquad \text{for } i = 1, \ldots, I \qquad \text{Eqn 7}$$

$$\underline{\underline{Y}}_j \quad \sim \quad \exp[\, N(\mu_j, \sigma_j) \,] \qquad \text{for } j = 1, \ldots, J \qquad \text{Eqn 8}$$

then the distribution of risk is also a LogNormal random variable of the form:

$$\underline{\underline{R}} \quad \sim \quad \exp[\, N(\mu_R, \sigma_R) \,] \qquad \text{Eqn 9}$$

with

$$\mu_R \quad = \quad \sum \mu_i - \sum \mu_j \qquad \text{Eqn 10}$$

$$\sigma_R \quad = \quad \text{Sqrt} \left[\sum \sigma^2_i + \sum \sigma^2_j \right] \qquad \text{Eqn 11}$$

with the sums over all the indicated indices. As discussed earlier, LogNormal distributions are self-replicating under multiplication and division.

3.2 Working with "High-End" and "Low-End" Values

In 1992, the US Environmental Protection Agency (US EPA) defined the concept of a "high-end" point value for a variable in the numerator of Eqn 2 as a deterministic input to an exposure assessment that falls above the 90th percentile but below the 99.9th percentile of the distribution for the particular random variable (US EPA, 1992). For a variable in the denominator of Eqn 2, one may define a corresponding "low-end" value as falling below the 10th percentile but not below the 0.1th percentile for the particular random variable.

For simplicity of exposition, let us take the 95th percentile as representing a high-end value and the 5th percentile as representing a low-end value of a distribution. Let the notations $\{\underline{\underline{T}}\}_{0.95}$ and $\{\underline{\underline{T}}\}_{0.05}$ and indicate the 95th and 5th percentiles, respectively, of an arbitrary random variable $\underline{\underline{T}}$.

With this notation, when the standard deviations are roughly similar, the high-end value of the numerator of Eqn 4 is considerably smaller than the function of the high-end inputs:

$$\{g(\underline{\underline{X}}_i)\}_{0.95} < g(\{\underline{\underline{X}}_i\}_{0.95}) \quad \text{for } i = 1, \ldots, I \qquad \text{Eqn 12}$$

Similarly, when the standard deviations are roughly similar, the low-end value of the denominator of Eqn 4 is considerably larger than the function of the low-end inputs:

$$\{h(\underline{\underline{Y}}_j)\}_{0.05} > h(\{\underline{\underline{Y}}_j\}_{0.05}) \quad \text{for } j = 1, \ldots, J \qquad \text{Eqn 13}$$

Overall, this means that the high-end value for risk is much, much smaller than the function of the high-end inputs in the numerator and the low-end inputs in the denominator when the standard deviations are roughly similar:

$$\{\underline{\underline{R}}\}_{0.95} \quad << \quad f(\{\underline{\underline{X_i}}\}_{0.95}, \{\underline{\underline{Y_j}}\}_{0.05}) \qquad \text{Eqn 14}$$

for i = 1, ..., I and j = 1, ..., J

Most risk assessors now understand this well-documented property of the fundamental risk equation, Eqn 2 (Burmaster & Harris, 1993; Bogen, 1994; Cullen, 1994). This property of the fundamental risk equation does not depend on the use of LogNormal distributions as inputs.

3.2 Working with Arithmetic Means

Let the notation $<\underline{\underline{T}}>$ indicate the arithmetic mean (or expected value) of an arbitrary random variable $\underline{\underline{T}}$. For a LogNormal distribution, the arithmetic mean is always greater than the median of the distribution by the factor $\exp[\, 0.5 \bullet \sigma^2_T \,]$. In many practical cases, the arithmetic mean of a LogNormal random variable falls between the 65th and the 80th percentiles of the distribution. However, in certain situations, the arithmetic mean of a LogNormal distribution can exceed the 95th percentile of that distribution.

Some mathematical properties hold in this situation. For independent LogNormal distributions, the arithmetic average of the numerator in Eqn 4 equals the function of the arithmetic averages of the input variables:

$$<g(\underline{\underline{X_i}})> \quad = \quad g(<\underline{\underline{X_i}}>) \qquad \text{for } i = 1, \ldots, I \qquad \text{Eqn 15}$$

Similarly, for independent LogNormal distributions, the arithmetic average of the denominator in Eqn 4 equals the function of the arithmetic averages of the input variables:

$$<h(\underline{\underline{Y_j}})> \quad = \quad h(<\underline{\underline{Y_j}}>) \qquad \text{for } j = 1, \ldots, J \qquad \text{Eqn 16}$$

The results in Eqns 15 and 16 are easy to prove for independent LogNormal distributions, and the results hold generally for other independent random variables from other families of distributions. Some authors use this property as the definition of independence between two random variables.

However, for independent LogNormal distributions, the arithmetic average of risk does not equal the function of the averages of the inputs:

$$\langle \underline{\underline{R}} \rangle \quad \neq \quad f(\langle \underline{\underline{X}}_i \rangle, \langle \underline{\underline{Y}}_j \rangle) \qquad \text{Eqn 17}$$

$$\text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J$$

This result in Eqn 17 surprises many people, even though it is easily proved for independent LogNormal distributions. It is true for other families of distributions as well.

3.3 Working with Medians

Let the notation $\{\underline{\underline{T}}\}_{0.50}$ indicate the median or 50th percentile of an arbitrary random variable $\underline{\underline{T}}$.

Some mathematical properties hold in this situation. For LogNormal distributions, the median of the numerator in Eqn 4 equals the function of the medians of the input variables:

$$\{g(\underline{\underline{X}}_i)\}_{0.50} \quad = \quad g(\{\underline{\underline{X}}_i\}_{0.50}) \qquad \text{for } i = 1, \ldots, I \qquad \text{Eqn 18}$$

and, the median of the denominator in Eqn 4 equals the function of the medians of the input variables:

$$\{h(\underline{\underline{Y}}_j)\}_{0.50} \quad = \quad h(\{\underline{\underline{Y}}_j\}_{0.50}) \qquad \text{for } j = 1, \ldots, J \qquad \text{Eqn 19}$$

More generally for independent LogNormal distributions, the median risk equals the function of the median inputs to Eqn 3:

$$\{\underline{\underline{R}}\}_{0.50} \quad = \quad f(\{\underline{\underline{X}}_i\}_{0.50}, \{\underline{\underline{Y}}_j\}_{0.50}) \qquad \text{Eqn 20}$$

$$\text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J$$

Thus, for independent LogNormal distributions, the median of the function for risk (in Eqns 2, 3, and 4) is the function of the median inputs. Although this result is

not true for independent random variables from other families of distributions, we have found it an excellent approximation in many numerical simulations of Eqns 2, 3, and 4.

3.4 Working with Mixed Cases

If we continue to restrict ourselves to independent LogNormal random variables as the inputs to the fundamental risk equation, any of Eqns 2, 3, or 4, then:

- the median of the R is equal to the function of the medians of the inputs;
- the arithmetic mean of R is NOT equal to the function of the arithmetic means of the inputs; and
- the 95th percentile of R is much smaller than the function of (i) the 95th percentiles of all the inputs in the numerator and (ii) the 5th percentiles of all the inputs in the denominator.

Thus, as is exactly true for independent LogNormal distributions and as is approximately true for other independent random variables with longer tails to the right, medians (not averages) are "neutral" and "self replicating" when used as point value inputs to the fundamental risk equation, Eqn 2.

Without doing a full calculation or a full simulation, no one can know the percentile of R calculated if the inputs to the fundamental risk equation, Eqn 2, include a combination of median values, average values, and high- and low-end values.

Restricting ourselves to the case with independent LogNormal distributions, we see that:

- the use of one or more median values in either the numerator or the denominator does not shift the estimate of R (further) above or (further) below the correct median of R, i.e., median inputs are "neutral" in trying to understand where the value R falls as a percentile of the distribution R;

- the use of one or more average values in the numerator does shift the estimate of R above the correct median of $\underline{\underline{R}}$, i.e., average inputs in the numerator introduce moderate to large (but unknown) amounts of conservatism in trying to understand where the value R falls as a percentile of the distribution $\underline{\underline{R}}$;

- the use of one or more high-end values in the numerator does shift the estimate of R far above the correct median of $\underline{\underline{R}}$, i.e., high-end inputs in the numerator introduce large (but unknown) amounts of conservatism in trying to understand where the value R falls as a percentile of the distribution $\underline{\underline{R}}$; and

- the use of one or more low-end values in the denominator does shift the estimate of R far above the correct median of $\underline{\underline{R}}$, i.e., low-end inputs in the denominator also introduce large (but unknown) amounts of conservatism in trying to understand where the value R falls as a percentile of the distribution $\underline{\underline{R}}$.

Most risk assessors now understand that the introduction of a few high-end values into the numerator or a few low-end values into the denominator of Eqns 1 or 2 can introduce very large amounts of conservatism into the point estimate R (Harris & Burmaster, 1992; Burmaster & Harris, 1993; Bogen, 1994; Cullen, 1994).

Fewer people understand that the introduction of several average values in the numerator of Eqns 1 or 2 can introduce significant amounts -- or even very large amounts -- of conservatism into point estimate R. As an extreme example, if the arithmetic means of three distributions all exceed the 90^{th} percentile of the corresponding distribution, the result is obvious. Less obvious, the use of three average values as point values for the corresponding LogNormal random variables can really be the multiplication of three 75^{th} percentiles. If these are the only conservative inputs in an equation, these three inputs may multiply to give, in effect, a high-end point value for risk. If these three average values for inputs in the numerator are combined multiplicatively with three high-end values for other inputs in the numerator, the resulting point estimate of risk may be far, far

more conservative than understood just from the combination of the three high-end values along with medians for the other variables.

4.0 Conclusions

From this discussion, we draw three main conclusions.

First, without doing a full calculation or a full simulation, no one can know the percentile of $\underline{\underline{R}}$ calculated if the inputs to the fundamental risk equation, Eqn 2, include a combination of median values, average values, and "high end" values.

Second, for independent LogNormal random variables -- and for other independent random variables from other families of distributions with long tails to the right -- the use of one or more medians in the numerator or denominator of Eqns 2, 3, or 4 for input variables does not introduce any compounding conservatisms; in contrast, the use of one or more average values in the numerator of those same equations always introduces multiplicative conservatisms, usually hidden from view and sometimes numerically large.

Third, the simultaneous use of several average values in the numerator (for distributions with long tails to the right) along with several high-end values in the numerator and several low-end values in the denominator can lead to point estimates of risk that fall above the range US EPA uses to set policy.

Acknowledgments

General Electric and Alceon supported this research.

References

Aitchison & Brown, 1957
Aitchison, J. and J.A.C. Brown, 1957, The Lognormal Distribution, Cambridge University Press, Cambridge, UK

Benjamin & Cornell, 1970
Benjamin, J.R. and C.A. Cornell, 1970, Probability, Statistics, and Decisions for Civil Engineers, McGraw Hill, New York, NY

Bogen, 1994
Bogen, K.T., 1994, A Note on Compounded Conservatism, Risk Analysis, Volume 14, Number 4, pp 379 - 381

Burmaster & Thompson, 1995, Need
Burmaster, D.E. and K.M. Thompson, 1995, The Need for New Methods to BackCalculate Soil CleanUp Targets in Interval and Probabilistic Risk Assessments, Human and Ecological Risk Assessments, in press

Burmaster & Thompson, 1995, BackCalculating
Burmaster, D.E. and K.M. Thompson, 1995, BackCalculating CleanUp Targets in Probabilistic Risk Assessments When the Acceptability of Cancer Risk is Defined Under Different Risk Management Policies, Human and Ecological Risk Assessments, in press

Burmaster & Harris, 1993, The Magnitude of Compounding Conservatisms in Superfund Risk Assessments, Risk Analysis, Volume 13, Number 2, pp 131 - 143

Crow & Shimizu, 1988
Crow, E.L. and K. Shimizu, Eds., 1988, Lognormal Distributions, Theory and Applications, Marcel Dekker, New York, NY

Cullen, 1994
Cullen, A.C., 1994, Measures of Compounding Conservatism in Probabilistic Risk Assessment, Risk Analysis, Volume 14, Number 4, pp 389 - 393

Evans et al, 1993
Evans, M., N. Hastings, and B. Peacock, 1993, Statistical Distributions, Second Edition, John Wiley & Sons, New York, NY

Feller, 1968 & 1971
Feller, W., 1968 and 1971, An Introduction to Probability Theory and Its Applications, Volumes I and II, John Wiley, New York, NY

Hattis & Burmaster, 1994
Hattis, D.B. and D.E. Burmaster, 1994, Assessment of Variability and Uncertainty Distributions for Practical Risk Assessments, Risk Analysis, Volume 14, Number 5, pp 713 - 730

Morgan, 1984
Morgan, J.T.M., 1984, Elements of Simulation, Chapman and Hall, London, UK

Ott, 1995
Ott, W.R., 1995, Environmental Statistics and Data Analysis, Lewis Publishers, Boca Raton, FL

Ott, 1990
Ott, W.R., 1990, A Physical Explanation of the Lognormality of Pollutant Concentrations, Journal of the Air and Waste Management Association, Volume 40, pp 1378 et seq.

Rubinstein, 1981
Rubinstein, R.Y., 1981, Simulation and the Monte Carlo Method, John Wiley & Sons, New York, NY

Springer, 1979
Springer, M.D., 1979, The Algebra of Random Variables, John Wiley & Sons, New York, NY

US EPA, 1992, Exposure
US Environmental Protection Agency, Guidelines for Exposure Assessment, 57 Federal Register, pp 22888 et seq., 29 May 1992