

A Simple Formula for Calculating the "Mass Density" of a Lognormally Distributed Characteristic: Applications to Risk Analysis

Adam M. Finkel¹

Received August 31, 1989; revised February 12, 1990

Statements such as "80% of the employees do 20% of the work" or "the richest 1% of society controls 10% of its assets" are commonly used to describe the distribution or concentration of a variable characteristic within a population. Analogous statements can be constructed to reflect the relationship between probability and concentration for unvarying quantities surrounded by uncertainty. Both kinds of statements represent specific usages of a general relationship, the "mass density function," that is not widely exploited in risk analysis and management. This paper derives a simple formula for the mass density function when the uncertainty and/or the variability in a quantity is lognormally distributed; the formula gives the risk analyst an exact, "back-of-the-envelope" method for determining the fraction of the total amount of a quantity contained within any portion of its distribution. For example, if exposures to a toxicant are lognormally distributed with $\sigma_{\ln x} = 2$, 50% of all the exposure is borne by the 2.3% of persons most heavily exposed. Implications of this formula for various issues in risk assessment are explored, including: (1) the marginal benefits of risk reduction; (2) distributional equity and risk perception; (3) accurate confidence intervals for the population mean when a limited set of data is available; (4) the possible biases introduced by the uncritical assumption that extreme "outliers" exist; and (5) the calculation of the value of new information.

KEY WORDS: Uncertainty; distribution of mass; Lorenz curve; risk perception; value of information.

1. INTRODUCTION

This paper introduces the concept of the "mass" of a quantity surrounded by uncertainty and/or variability, and attempts to show how bringing to bear information about "mass" may lead to changes in how we assess, communicate, and control environmental and health risks. The formulas derived herein will facilitate such improvements, especially when the uncertainty or variability can be modeled using a lognormal probability density function (PDF). Figure 1 shows a discretized lognormal PDF for the varying quantity "annual in-

come" in a hypothetical population of 19,935 persons, who together earn nearly \$253 million each year. A contemporary "risk analysis" would probably generate one or more point estimates to describe this situation; for example, that the median income is \$10,000, that the mean is \$12,690, or that the "plausible upper bound" (95th percentile) is approximately \$28,000. The numbers above the histogram in Fig. 1 represent the total amount of money earned by persons at each income level—in other words, the "mass" of each category. In the discretized case, these values lead directly to additional and currently untapped information about the concentration of the varying characteristic—for instance, that "the 3.6% of persons earning between \$40,000 and \$60,000 control 12.5% of the total wealth." The remainder of this paper

¹ Resources for the Future, Center for Risk Management, 1616 P Street, N.W., Washington, D.C. 20036.

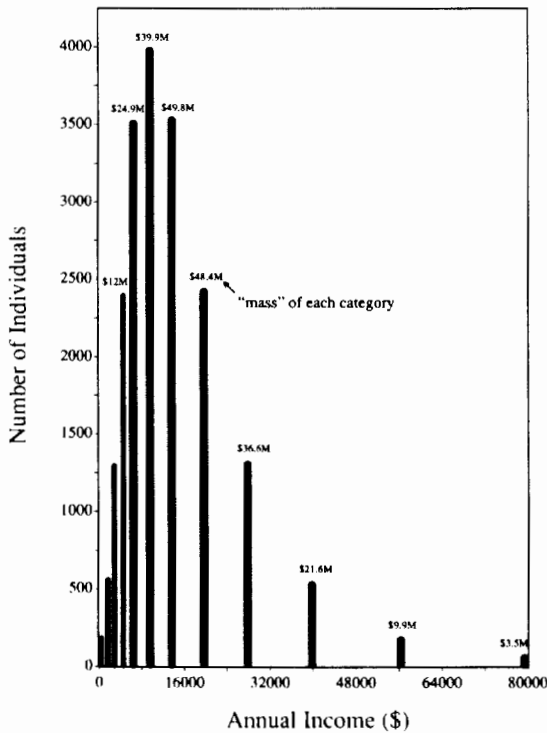


Fig. 1. Discretized version of a lognormal PDF, describing the variability of annual income in a hypothetical population. The numbers above the vertical bars represent the product of the number of persons at each income level and their annual income—that is, the “mass” of each category.

will explore the “mass” of characteristics that can be described as continuously distributed (specifically as lognormal), using examples relevant to health risk assessment.

1.1. Mathematical Preliminaries

Suppose x represents a nonnegative physical characteristic that varies among a population of T individuals, and that T is large enough that we can use a continuous PDF $f(x)$ to model the (discrete) values of x and their associated probabilities. Alternatively, $f(x)$ might represent a degree-of-belief or uncertainty distribution about the true value of an invariant physical characteristic (or some other quantity such as a rate, proportion, or risk). In the first case, $T \int_a^b f(x) dx$ is a measure of the number of individuals for whom $a \leq x < b$; in the second case the integral $\int_a^b f(x) dx$ measures the subjective or objective probability that $a \leq x < b$. Just as the integral $\int_a^b xf(x) dx$ yields the mathematical expectation of x , or $E(x)$, the

quantity $T \int_a^b xf(x) dx$ measures the total amount of the characteristic contained within the population.

When integrals containing the expression $xf(x) dx$ are evaluated over limits not spanning the entire domain of x , they provide information about the *concentration* of the characteristic within the appropriate portion of the PDF. One can compare integrals of $f(x)$ between two sets of limits and discern how the “amount of probability” depends on x ; similarly, one can compare definite integrals of $xf(x)$ and discern how the “amount of mass” depends on x . In the same sense that the PDF $f(x)$ measures the amount of probability as a function of x , the “mass density function” $xf(x)$ measures the amount of (probability times magnitude) as a function of x . The ideas of mass and concentration find their way into everyday usage via rules-of-thumb like the “80/20 rule,” which can be invoked to explain that the most productive 20% of workers in a company accomplish 80% of all the work.

More formally, the “mass distribution function” $g(a, b)$ can be defined as:

$$g(a, b) = \frac{\int_a^b xf(x) dx}{\int_0^\infty xf(x) dx} \quad 0 \leq a < b \quad (1)$$

The denominator, which is equal to $E(x)$, normalizes $g(a, b)$ such that its minimum and maximum are zero and unity. As a simple example, let $f(x)$ be the uniform PDF over the interval $[0, 12]$ —so $f(x) = 1/12$ at all points within the interval. Then $g(a, b) = (b^2 - a^2)/144$. If x was the hourly income of each individual in a population, $g(0, 6) = 0.25$ means that one quarter of all the income in the population accrues to persons earning at or below the median wage. If x was instead the uncertain value of a particular individual’s wage, $g(11, 12) = 0.1597$ implies that the estimate $E(x) = 6$ is highly sensitive to the chance that x is between 11 and 12, because nearly one-sixth of the total mass is contained within the topmost one-twelfth of the distribution. This relationship between the mass within a portion of the PDF and the probability within that portion is formalized via the “concentration function”⁽¹⁾:

$$h(a, b) = \frac{\int_a^b xf(x) dx}{\int_a^b f(x) dx} \quad (2)$$

By definition, $h(a, b)$ is also equal to the expected value of x within the truncated portion of its domain between a and b . For example, $h(11, 12)$ for the uniform PDF equals $0.9583 \div (1/12)$, or 11.5.

For the special case when $a = 0$, $G(b) = E(x)^{-1} \cdot \int_0^b xf(x) dx$ is the (normalized) cumulative mass distribution function. $G(b)$ is analogous to the cumulative prob-

ability density function, or CDF, $F(b) = \int_0^b f(y) dy$.² Solving for x in the equation $F(x) = 0.5$ gives the median of the PDF, the point above and below which half of all the values lie. Similarly, the value of b which solves $G(b) = 0.5$ can be termed the “center of mass,” in that half of the total quantity of the characteristic lies above and below this point. If, for example, a metal bar 12 inches long and 1 inch square (representing a uniform PDF over the interval $[0,12]$) was constructed of an alloy of varying composition whose physical density was equal to x g/in³ for all values of x , the bar would weigh 72 g (12 in³ times the average density of 6 g/in³). It could be balanced on a fulcrum located to the right of the center of the bar, at the point $x = \sqrt{72}$, because $\int_0^{\sqrt{72}} x(1/12) dx \div 6 = 0.5$. In general, the center of mass \hat{m} for a uniform distribution over the interval $[a, b]$ is given by: $\hat{m} = \sqrt{(a^2 + b^2)/2}$.

For all other PDFs that contain more information than the uniform distribution, an analytic expression for $g(a, b)$ or $h(a, b)$ may be quite complicated or impossible to obtain, despite the usefulness of these functions. This paper derives a simple equation for the mass distribution function of the lognormal PDF, one of the distributions used most frequently in health and environmental risk analysis (as well as in many other fields). The applications discussed below may be sufficiently important to warrant derivation of the mass distribution function via numerical methods for other PDFs that arise in particular risk management cases. This result was derived during the preparation of the author’s doctoral dissertation⁽³⁾; it arose independently of the related result reported by Quensel⁽⁴⁾ in his study of the moments of truncated lognormal distributions.

2. DEFINITIONS

- X = lognormally distributed random variable
- \hat{X} = median (geometric mean) value of X ($\ln \hat{X}$ is the arithmetic mean of the normal variable $\ln X$)
- $E(X)$ = arithmetic mean value of X
- σ = standard deviation (SD) of $\ln X$ (note: the “geometric SD,” commonly written σ_g , equals e^σ)
- $f(X)$ (lognormal PDF for X) = $\frac{1}{\sigma X \sqrt{2\pi}} \cdot \exp \frac{-(\ln X - \ln \hat{X})^2}{2 \sigma^2}$

N_X = unit normal transformation; $N_X = \ln(X/\hat{X})/\sigma$
 therefore, $X = \hat{X} \exp(N_X \sigma)$
 $dX = \sigma \hat{X} \exp(N_X \sigma) dN_X$
 (note: for typographical convenience, “ N ” will sometimes be used in place of “ N_X ”)

$\phi(N)$ = the unit normal PDF = $\frac{1}{\sqrt{2\pi}} \cdot \exp(-0.5 N^2)$

$\Phi(y)$ = the unit normal CDF = $\int_{-\infty}^y \phi(N) dN$

3. DERIVATION

The unit normal transformation gives

$$\int_a^b X f(X) dX = \int_{N_a}^{N_b} \frac{\hat{X} e^{N\sigma} \exp(-0.5N^2) (\sigma \hat{X} e^{N\sigma}) dN}{\sigma(\hat{X} e^{N\sigma}) \sqrt{2\pi}} \quad (3)$$

$$= \hat{X} \int_{N_a}^{N_b} \frac{1}{\sqrt{2\pi}} \exp(N\sigma - 0.5N^2) dN \quad (4)$$

By completing the square in the integrand:

$$= \hat{X} \int_{N_a}^{N_b} \frac{1}{\sqrt{2\pi}} \exp(0.5 \sigma^2) \exp(-0.5(N - \sigma)^2) dN \quad (5)$$

Let $Z = (N - \sigma)$; $dZ = dN$; note that $E(X) \equiv \hat{X} \exp(0.5\sigma^2)$

$$= E(X) \int_{N_a - \sigma}^{N_b - \sigma} \frac{1}{\sqrt{2\pi}} \exp(-0.5Z^2) dZ \quad (6)$$

4. RESULTS

Thus, for the lognormal PDF,

$$g(a, b) = \int_{N_a - \sigma}^{N_b - \sigma} \phi(Z) dZ = \Phi(N_b - \sigma) - \Phi(N_a - \sigma) \quad (7)$$

In other words, *the fraction of the total mass of the lognormal quantity that is contributed by values between a and b equals the area under the corresponding unit normal PDF between $(N_a - \sigma)$ and $(N_b - \sigma)$.*

Recall that $N_y = +3$ is equivalent to $(\ln y = \ln \hat{X} + 3\sigma)$ or $(y = \hat{X} e^{3\sigma})$; semantically, we will refer to y as being “3 SD above” (or “to the right of”) the median. For example, consider a lognormal quantity with $\sigma = 2$ (i.e., an uncertainty or variability of about a “factor of 50” in either direction).³ By Eq. (7), the mass contributed by values between -3 and $+3$ SD from the

² Aitchison and Brown⁽²⁾ discuss $G(b)$ in passing, referring to it as the “first-moment distribution function.”

³ We will refer to the quantity $\{\exp(1.96\sigma)\}$ as the “uncertainty factor” (UF), since 95% of the probability density lies between \hat{x}/UF and $\hat{x} \cdot UF$.

median equals the unit normal area between (-5) and $(+1)$, which is approximately 84%. Figure 2 plots $G(b)$ against $F(b)$, thereby forming the "Lorenz diagram" familiar to economists, for four different lognormal distributions. Notice, for example, that for $\sigma=2$ the first 80% of the population contributes about 12.5% of the mass; $F(b) = 0.8$ involves the portion of the population up to 0.85 SD above the median [$\Phi^{-1}(0.8)=0.85$ and $\Phi(0.85-2)=0.125$].

Equation (7) has several obvious corollaries.

(1) *The cumulative fraction of mass to the right of K standard deviations above the median equals the cumulative unit normal area beyond $(K-\sigma)$.* Thus, for example, even though there is only about one chance in 30,000 that a value beyond 4 SD will occur, if $\sigma=2$ these values contribute 2.3% of the total mass (the area of the unit normal PDF beyond $+2$); if $\sigma=3$, these values would contribute 16% [$1-\Phi(4-3)$] of the total mass.

(2) *More specifically, the "center of mass" of the lognormal lies at exactly σ standard deviations above the median.* When $K=\sigma$, the fraction of total mass beyond K SD equals the unit normal area to the right of $(\sigma-\sigma)$, or zero; by definition, this fraction equals one-half. Therefore, $N_m=\sigma$, or $\hat{m}=\hat{X}\exp(\sigma^2)$. For highly skewed lognormals, this rule-of-thumb provides a stark reminder of the influence of "outliers." For example,

if $\sigma=2$ (UF=50.4), fully half the mass comes from the 2.3% of values more than 2 SD above the median (i.e., more than 54.6 times the median). Note that the center of mass is always to the right of the mean; because $E(X) \equiv \hat{X}\exp(0.5\sigma^2)$, $\hat{m}/E(X)=E(X)/\hat{X}$.

(3) *The "complementarity point" is the point on the PDF where $C\%$ of the values contribute $(100-C)\%$ of the mass, and vice versa; the value of C is found via the unit normal transform of $(\sigma/2)$.*⁴ For example, if $\sigma/2=1$, 84% ($1-\Phi(+1)$) of the mass is concentrated in the largest 16% of values. Table I shows the value of C at the complementarity point, for various values of σ . For instance, the "80/20" rule applies to a lognormal quantity with $\sigma=1.69$ (UF=27.5).

Note also that because $N_{E(X)}=\sigma/2$, the mean and the complementarity points are coincident; graphically, these points can be found where the Lorenz curve crosses the diagonal line $y=1-x$. Figure 3 shows the percentile (p) of the distribution of X at which the mean is located, as a function of σ ; for σ greater than about 0.7, the log-linear regression equation $p=(69.15 + 21.41 \ln \sigma)$ yields a good approximation ($R^2>0.99$) for p . For sufficiently skewed distributions ($\sigma>3.29$, UF>631.7), the mean exceeds the upper 95th percentile of the distribution, casting into question whether this probabilistic "upper bound" is in fact "conservative" in an expected-value sense.⁽⁵⁾

5. APPLICATIONS TO RISK ANALYSIS

As was discussed in the introduction, if the PDF reflects variability across a population, $g(a,b)$ is a measure of the total amount of the characteristic within the population or a portion thereof. In environmental and occupational health risk analysis, variability in exposure

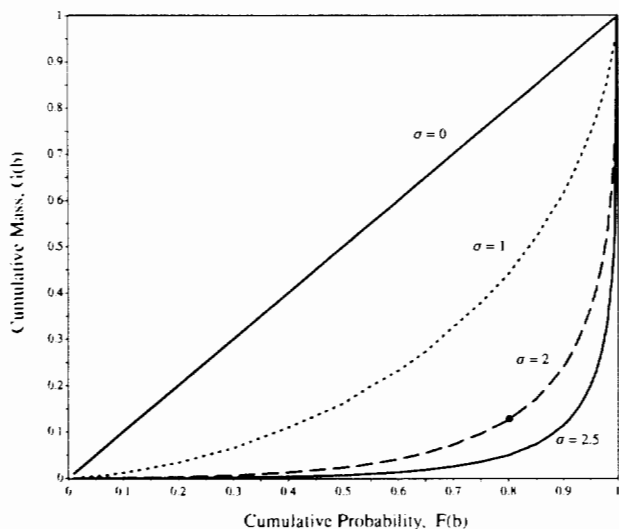


Fig. 2. "Lorenz curves" for four lognormal PDFs of varying breadth. The point indicated on the Lorenz curve for $\sigma=2$ (UF=50.4) reveals that the first 80% of the population accounts for only 12.5% of the "mass;" the remaining 87.5% of the mass is concentrated among the remaining 20% of the population.

Table I. Concentration of $C\%$ of the Mass in $(100-C)\%$ of Values

σ	Division of probability:mass at "complementarity"
0.5	60:40
1.0	69:31
1.69	80:20
2.0	84:16
2.5	89:11
3.0	93:7
3.5	96:4

⁴ This corollary follows from solving the equation $\Phi(W-\sigma)=1-\Phi(W)$, invoking the identity $\Phi(-x)=1-\Phi(x)$; the solution is $W=\sigma/2$.

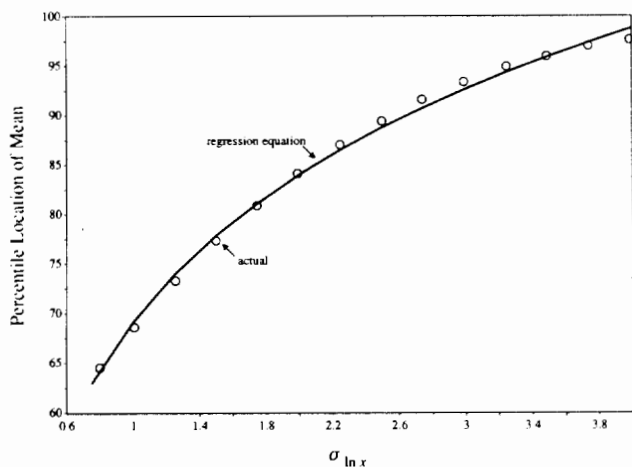


Fig. 3. Location of the mean of a lognormal distribution (the percentile of the CDF coincident with the mean) as a function of σ , showing that the mean will exceed any arbitrary probabilistic "upper bound" as the uncertainty or variability increases. The open circles represent the exact values of $100 \cdot \Phi(\sigma/2)$. The solid line indicates the best-fitting log-linear regression equation in the range $0.7 < \sigma < 4.0$.

is commonplace, and is often modeled using lognormal distributions.⁽⁶⁾ In addition, a recent study of interindividual variation in human susceptibility to carcinogenic stimuli⁽⁷⁾ concluded that variability in the effective carcinogenic potency of a substance may plausibly be described as lognormal, and may be rather broadly distributed ($UF \approx 100$).

If, on the other hand, the PDF reflects uncertainty about the true value of an invariant quantity, $g(a, b)$ measures the extent to which different portions of the PDF affect the estimate of the expected value of the quantity. Various researchers have generated, via statistical analyses, simulation modeling, or elicitation of expert opinion, PDFs for such uncertain quantities as the downgradient pollutant concentration available for inhalation⁽⁸⁾ or ingestion,⁽⁹⁾ the biologic potency of carcinogens^(10,11) or neurotoxins,⁽¹²⁾ and the relationship between administered and delivered dose.⁽¹³⁾ Of course, risk-based decisions often involve both variability and uncertainty.⁽¹⁴⁾ In such cases, the regulator might be concerned, for example, with the fraction of total risk borne by highly exposed (or highly susceptible) individuals, conditional on the true potency of the substance in question being in the "tail" of its own uncertainty distribution. The individual confronted with the same combination of PDFs might well view his risk as the product of random or conservative "draws" from each distribution (How potent is the substance? How susceptible am I relative to the "average" person?).

Some of these issues can be addressed without reference to the mass distribution function. For example, Rappaport *et al.*⁽¹⁵⁾ showed that when occupational exposures are distributed lognormally (spatially and/or temporally), one can constrain the probability or frequency of "excursions" above an arbitrary threshold or standard by maintaining the mean concentration at or below a defined fraction of the standard. This decision-rule, however, does not address the question of how much of the total exposure or risk will be borne by the fraction of workers subject to the excursions.⁽¹⁶⁾

The ability to discern the relative contributions of different portions of a distribution to its total mass can potentially improve risk assessment, risk management, or risk communication in the following four areas.

5.1 Distributions and Marginal Distributions of Consequences

Information about the fraction of the total exposure or risk attributable to part of a distribution can elucidate properties of the total benefit and marginal benefit functions that would not be apparent if only the absolute values of risk were known. The following three examples will illustrate various situations in which costs or benefits depend implicitly on the mass distribution.

5.1.1. Benefits of Targeted Risk Control

One of the prime motivations for investigating the distribution of exposures or of human susceptibilities is to gauge whether risk reduction measures would be more effective if targeted at particular segments of the population, rather than implemented across-the-board. The mass distribution function provides a screening device to make such determinations. For example, Cohen and Gromicko⁽¹⁷⁾ investigated the distribution of indoor radon levels in many U.S. states, including about 5000 homes in Pennsylvania. In that state, the exposure distribution was well-described by a lognormal with $\bar{x} = 3$ picocuries per liter (pCi/l) and $\sigma = 1.28$. Using Eq. (7), therefore, one could estimate that the 5% of homes containing the highest radon levels (i.e., those above 24.6 pCi/l) account for about 36% [$1 - \Phi(1.645 - 1.28)$] of the total cancer risk from radon in this sample, assuming that risk is linear in exposure.⁵ Equivalently, by targeting

⁵ Note that Eq. (2) and (7) together also provide a rapid method for assessing the average exposure (AE) faced by persons in any portion of the exposure distribution; for the topmost 5%, $AE = [0.36 \cdot E(x)] - 0.05 = 49.0$ pCi/L.

exposure mitigation efforts at houses containing more than 24.6 pCi/l, society might be able to “solve” more than one-third of the problem at perhaps one-twentieth of the expense of a more sweeping program.

Puskin and Nelson⁽¹⁸⁾ recently analyzed similar data (using a lognormal distribution for indoor radon exposure with a median of 0.9 pCi/l and $\sigma = 1.16$), and performed their own calculations of the relationship between probability and mass. For these calculations, they developed a computer program to numerically integrate the mass density function, rather than using a simple relationship of the type presented in this paper (Puskin and Nelson, personal communication). They acknowledged that targeting nationwide mitigation efforts at the approximately 600,000 U.S. homes that have radon levels above 10 pCi/l might be most cost-effective; however, they also made it clear that in order to reduce the total number of deaths attributable to radon by more than 17%, reductions must be made in homes with lower levels as well.

Similarly, Schwing and Kamerud⁽¹⁹⁾ determined that the risk of death to a motor vehicle occupant (per person-mile driven) varies substantially with the time of day and the day of the week, such that the risk in the most dangerous single hour of the week (4.3×10^{-7} /mile, around midnight on Saturday/Sunday) is more than 134 times as risky as the least dangerous hour (3.2×10^{-9} /mile, around 8 a.m. Sunday). Figure 4 shows the Lorenz curve that Schwing and Kamerud constructed, along with an exact lognormal distribution ($\sigma = 0.8$), which provides a reasonable approximation to the empirical data.

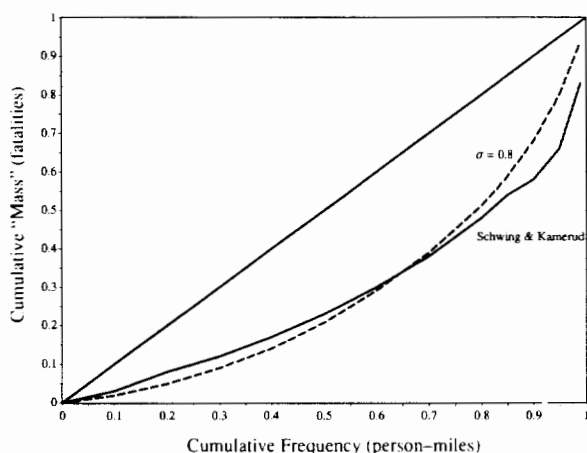


Fig. 4. Lorenz curves demonstrating the concentration of traffic fatalities during certain hours of the week. The solid line is derived from the empirical data analyzed by Schwing and Kamerud.⁽¹⁹⁾ The dashed line is the best-fitting lognormal approximation ($\sigma = 0.8$).

They concluded that the optimal allocation of police activity might well depend on the time of day, citing as evidence the fact that the most dangerous 1% of all travel results in 17% of all fatalities. They also concluded that across-the-board controls (e.g., the national 55 mph speed limit) are likely to be less efficient than “more elaborate (i.e., disaggregated) policies,” although they imply that such policies may have their own drawbacks.

5.1.2. Standard-Setting and “Acceptable” Risk

Even in cases where risk managers cannot identify the most susceptible or most highly exposed individuals, or where no control measures exist to preferentially reduce some individual risks, they may wish to examine the mass concentration of risk, in the context of “acceptable” individual risk levels. Although by reducing risks across-the-board, the regulator cannot influence the shape of the interindividual distribution of risk, he can influence both the number of persons whose risk levels exceed some “bright line” and the absolute number and the relative fraction of deaths expected to occur among these individuals. Depending on the parameters of the risk distribution, their relation to the “bright line,” and the implicit value judgments made in managing the risk, the marginal benefit function for across-the-board mitigation of risk may demonstrate important nonlinearities.

For example, suppose that in a population of 10 million people, a particular excess cancer risk is lognormally distributed ($\sigma = 2$) about a median value of 1.35×10^{-6} . This value was chosen so that the mean risk equals 10^{-5} , thus yielding an expected excess death toll of 1000. Suppose further that the cutoff of “acceptable” risk is set at 10^{-4} . Then in the baseline situation, 1.6% of the population would face “unacceptable” risks, but 44% of the 1000 expected excess deaths would occur in this subpopulation. A 10-fold reduction in all exposures would reduce the number of persons “above the line” by a factor of 32 [10^{-4} moves from being 2.15 SD above the original median to being 3.30 SD above the new median of 1.35×10^{-7}]. At the same time, however, the number of deaths in this subpopulation will fall by a factor of 44 (from 44% of 1000 total deaths to 10% [$1 - \Phi(3.30 - 2)$] of 100 total deaths).

Table II shows other values for the partitioning of expected deaths among persons on either side of the “bright line.” As the public continues to become more aware of the importance of individual-risk cutoffs in regulatory decision-making and the impossibility of ensuring that all persons face zero or even *de minimis* risks, the pressure may mount for risk managers to consider

Table II. Hypothetical Risk Management Case Showing Nonlinear Benefit Functions

Exposure reduction factor	Total deaths expected	% of population in "unacceptable" region (above $R = 10^{-4}$)	% of deaths in "unacceptable" region	No. of deaths in "unacceptable" region
1 (baseline)	1000	1.6	44	440
2	500	0.7	31	155
5	200	0.2	17	34
10	100	0.05	10	10
50	20	0.002	1.8	0.36
100	10	0.0005	0.7	0.07

the absolute and relative consequences to those whose exposures and/or susceptibilities leave them "above the line." The time may now be ripe for study of how the ineluctable tradeoff between "deaths among the majority" and "deaths in the tail" adds connotations to the larger tradeoff (also visible in Table II) between total deaths and total cost.

Interestingly, welfare economists have long realized that measures of national poverty which consider only the "head count" of persons below the "poverty line" are insensitive and misleading.⁽²⁰⁾ Sen, for example, implicitly concluded that both the PDF and the mass distribution function were essential inputs to any truly useful measure of poverty; his "poverty index" is a function of the average dollar shortfall a poor person faces relative to the poverty line and of the extent of distributional inequality, as well as of the "head count" itself.⁶ The analogy between income shortfall and the exceedance of individual risks above "acceptability" is a potent one, and the role of large equity differences between the fortunate and unfortunate may be similar in both contexts.

5.1.3. Distributional Equity of Variable Risks

The "average risk" and the "maximum plausible risk" are useful decision tools, irrespective of whether the PDF for risk is a consequence of uncertainty, variability, or both. However, both measures assume rather

specialized utility or "willingness-to-pay" functions to weigh the costs of different possible outcomes. The average risk $E(X)$ [whether expressed as a probability or as the "body count" $T \cdot E(X)$] implies that social cost is linear in risk or incidence; use of the upper-bound value implies a sharp discontinuity in the cost function associated with a "bright line" of acceptable or *de minimis* risk. An ideal measure of expected social cost (alternatively, the expected benefit of eliminating the risk) would account for the possible nonlinearity of individual utility functions (for uncertain risks) and/or of social welfare functions for aggregating individual costs (for variable risks).⁽²¹⁾ Such a measure would require the evaluation of the integral $\int_R d(R)f(R)dr$, where $d(R)$ is the "damage function" over risk and $f(R)$ is the PDF for risk. Again, this is analogous to the welfare economics research of Sen,⁽²⁰⁾ Atkinson,⁽²²⁾ and others, who generally suggest that not only is individual poverty a function of income shortfall, but that that function should change more steeply than linearly as the shortfall increases. In the context of risk, specifying the damage function is a daunting task, because of the controversial value judgments needed and the analytic complexity of the required calculations.

However, appreciation of the mass distribution can lead to a middle ground between oversimplification and unmanageable complexity. For example, consider the possibility that human susceptibilities to a given carcinogen are lognormally distributed with $UF \approx 100$ ($\sigma \approx 2.35$). This implies that in the absence of any systematic relationship between exposure and susceptibility, the least susceptible 10% of the population bears only about 0.02% [$\Phi(-1.3 - 2.35)$] of the population risk, while the most susceptible 10% bears about 85% of the risk. Lopes⁽²³⁾ has shown empirically that when confronted with various hypothetical monetary lotteries, where both the expected value and the variance remains fixed, people generally

⁶ More specifically, the poverty index depends in part on the "Gini coefficient," G , a measure of the overall departure from distributional equality. G , which varies between 0 and 1, equals twice the area between the Lorenz curve and the 45° line of perfect distributional equality. For reference, when $\sigma = 1$, $G = 0.52$; when $\sigma = 2$, $G = 0.84$.⁽²⁾

react unfavorably to the hypothetical gambles in direct proportion to the amount of inequality between the “mass” of the upper and lower portions of the PDF. This phenomenon may reflect the individual’s feelings of the injustice of situations where costs and benefits are highly concentrated, and/or personal risk aversion to the possibility of large risks that becomes more pronounced as the skewness increases. In any event, risk managers may better understand the “disparities between the subjective risk opinions of the lay public and the objective risk calculations of experts”⁽¹⁹⁾ if they evaluate uncertain risks *vis-a-vis* their concentration as well as their magnitude.

5.2. Errors in Assessed Means/Variances from Limited Sample Data

Because the mass distribution function for an uncertain quantity can illustrate how values in different portions of the PDF influence the estimate of $E(X)$, it can shed light on a problem in statistical inference that frequently arises (but is often ignored) in risk analysis—determining the appropriate confidence bounds on the mean of a lognormally distributed parameter with limited observational data. Risk analysts may be familiar with the central limit theorem, which states that as the number n of independent observations (each from a distribution with common variance V) approaches infinity, the distribution of the sample mean of those observations can be approximated by a normal distribution with variance (V/n) . They may also recall the rule-of-thumb which deems this approximation valid in practice when n is greater than 30.

However, this rule-of-thumb assumes that the underlying population from which the observations come (the “parent” distribution) is itself reasonably symmetric (e.g., Gaussian). For highly skewed parent distributions like the lognormal, the central limit theorem’s approximation may be very misleading even when n is “large.” Only a few papers, mostly in the engineering literature,^(24,25) have investigated the distribution of the sum (and hence the mean) of n independent lognormal variates. Although the exact expression for this PDF remains analytically intractable, these researchers have concluded that it is approximately *lognormal* with a logarithmic SD σ^* inversely proportional to n (a result not inconsistent with the central limit theorem, since as $\sigma \rightarrow 0$ the lognormal becomes indistinguishable from the corresponding normal distribution).

I have validated this conclusion via Monte Carlo simulation. Figure 5 shows σ^* for sets of 5000 simulated observations, with each observation representing the mean

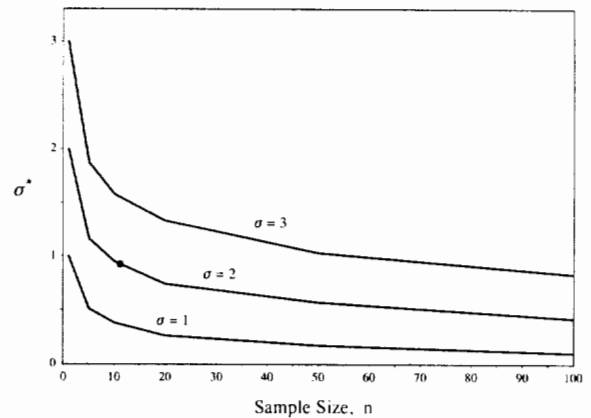


Fig. 5. A family of curves demonstrating the relationship between the sample size n and the SD of the mean of the n observations, drawn from lognormal parent distributions of varying breadth. The curves connect data points representing the empirical SD of the natural logarithms of 5000 Monte Carlo simulations, each one consisting of n observations.

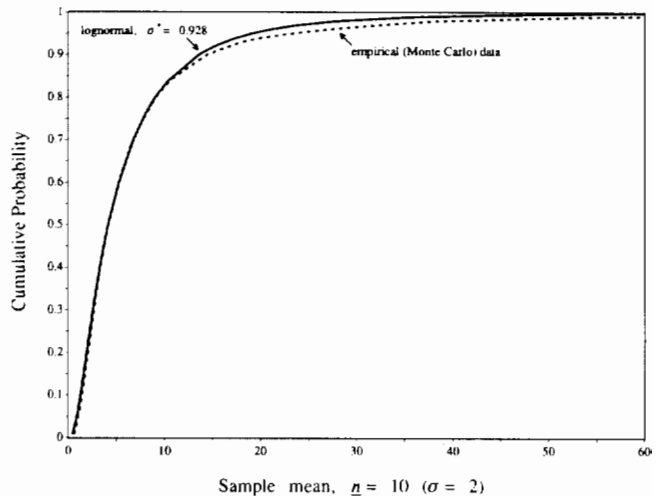


Fig. 6. The CDF of the Monte Carlo data for the data point indicated by the solid circle in Fig. 5. The dashed line is the empirical CDF derived by plotting the 5000 observations (each representing the mean of 10 samples) in ascending order. The solid line is the exact lognormal CDF which best approximates this data ($\hat{x} = 4.126$, $\sigma^* = 0.928$).

of a set of n data points drawn from parent distributions of various breadth (in all the examples that follow, $\hat{x} = 1$). Figure 6 shows the simulated data for one of these combinations ($\sigma = 2$, $n = 10$), along with the exact lognormal which best approximates this distribution. The Taylor expansion referenced in Crow and Shimizu⁽²⁶⁾ for the variance of the distribution of means yields a close ap-

proximation to the variance of the simulated data when σ is less than about 2.5; using the general formula for the variance of a lognormal $\{V = [E(X)]^2 \cdot (\exp(\sigma^2) - 1)\}$, one can then solve for σ^* . For very large σ , an approximation developed by Farley and referenced in Schwartz and Yeh⁽²⁵⁾ may yield more accurate results, or a nomogram such as that in Fig. 5 can be used.

When confronted with a limited set of sample data from a lognormal distribution, there are practical drawbacks to using the narrow and symmetric confidence intervals about the observed mean given by Gaussian statistics.⁽¹⁶⁾ Most significantly, because of the asymmetric uncertainty, the observed mean is more likely to be an underestimate of the true mean than an overestimate, but the absolute magnitude of any error of overestimation will be greater than an error of underestimation. For example, the Monte Carlo analysis in Fig. 5 shows that σ^* is approximately 0.93 ($UF = 6.4$) when $n = 10$ and the parent distribution has $\sigma = 2$. Knowing nothing about the true mean of the parent distribution but the 10 samples observed, the correct assumption is that while the observed mean is the best estimate of the true mean, the true mean is likely to be somewhat higher, but possibly very much lower.

The mass density function provides support for this somewhat counterintuitive phenomenon, as well as a handy way to gauge the breadth of the confidence intervals about the observed mean. For example, in a sample of 100 observations from a parent distribution with $\sigma = 2$, the binomial theorem shows there is a strong possibility (0.995^{100} , or about 60%) that none of the observations will come from the "tail" of the parent distribution exceeding 2.6 SD above its median ($x > 181$). By Eq. (7), this portion of the parent distribution contributes $[1 - \Phi(2.6 - 2)]$ or 27.5% of the mean, so there is a strong likelihood that underrepresentation of this portion will cause the observed mean to be as much as 27.5% lower than the true mean of 7.4. On the other hand, there is about 1% chance that 3 of the 100 observations will come from the "tail"—in such a case, the observed mean might be a factor of 2 or more above the true mean.⁷

5.3. Errors Due to Assuming an Exact Distribution/ Ignoring Truncation

The mass distribution function can also be edifying in situations that pose problems converse to those men-

tioned above. Frequently, instead of working with observed data and trying to infer the parameters of the distribution to which these data belong, the risk assessor has a hypothesis about a distribution that implies the existence of particular values of the quantity. In such cases, the "model uncertainty" (i.e., the assumption that information about probabilities and magnitudes can be obtained by applying formulas appropriate to an exact lognormal distribution) may contribute in subtle but substantial ways to the total uncertainty in the problem.

In practice, the temptation is to assume that the frequency or state-of-knowledge distribution of a quantity is lognormal, with a right-hand tail extending asymptotically to infinity, on the basis of observational data that "looks" lognormal (but that probably includes no extreme outliers) or physical principles that lead to lognormality (e.g., many independent and multiplicative sources of variability or error). However, physical, mathematical, or practical constraints may preclude the actual existence of such outliers (or of extreme values of the state-of-knowledge distribution). Even if one's decision process does not explicitly hinge on, or even consider, the magnitude of exposures or risks among extreme outliers, the implicit assumption that such outliers exist may lead to biases in one's assessment of critical decision variables, notably the mean. For example, Apostolakis and Kaplan⁽²⁷⁾ undertook a study of the reliability of engineered systems, and showed that the common assumption that the cumulative probability of failure is proportional to the failure rate λ can introduce severe biases in the assessed mean of the time-to-failure distribution. The bias occurs because of the use of (λt) as an approximation for $(1 - e^{-\lambda t})$ breaks down when λt approaches unity. For example, in a system with λ lognormal ($\hat{\lambda} = 6.75 \times 10^{-5}$, $\sigma = 2.52$), the assessed mean would be 1.62×10^{-3} if values of λt greater than unity were permitted; they showed that the true mean when $(1 - e^{-\lambda t})$ is used is only 1.08×10^{-3} (33% lower).

Similar situations where exponentially transformed or truncated lognormals may be more appropriate than exact lognormals also arise frequently in health risk analysis. In the workplace, airborne concentrations of a volatile toxicant may be lognormally distributed in the central portion of the distribution, but truncated at some higher value, either because of physical constraints (e.g., existence of a saturation vapor pressure for the substance) or practical realities (e.g., presence of an irritant effect or odor threshold causing workers to leave the area or suspend the production process). When considering uncertainties in biologic potency, in addition to the mathematical constraint that risks cannot exceed 1.0,

⁷ In addition, because the variance of a lognormal distribution is proportional to the square of the mean, estimates of the variance based on limited observational data are likely to be even more uncertain than corresponding estimates of the mean, as Apostolakis and Kaplan⁽²⁷⁾ observed.

there are other reasons to be suspicious of the tails of putative distributions of susceptibility. One might doubt, for example, that fetuses genetically constituted to be thousands or millions of times more susceptible to cancer than average would survive the gestation period—in any event, a risk manager might conclude that such “hyper-susceptibles” were destined to develop the disease at a high rate independent of anthropogenic stimuli or regulatory controls.

The mass distribution function informs the risk assessor how sensitive the mean is to valid or invalid parameter values in the tail of the PDF. In addition, examination of the mass distribution function may provide an impetus for research directed at reducing the overall uncertainty or ruling out the possibility that values in the tail need to be entertained, as it gives an indication of the “payoff” expected from reducing potential bias in the mean. For example, a lognormal uncertainty distribution for the potency of a carcinogen obeying the “90/10 rule” ($\sigma = 2.58$, $UF = 157$) has a mean which would be nine times smaller if values in the topmost 10% of the tail could be ruled out. Conceivably, a bioassay with more than the usual number of test animals could reduce the spread in the distribution due to random variation and model uncertainty (is the true dose-response function linear or quadratic?), or a well-designed epidemiology study might have sufficient power to exclude the potency values in the tail.

Two nuances of the sensitivity of the mean to the outliers are worth mentioning. First, when one believes that σ is large, it is worth investigating this sensitivity even if the possible truncation point (or the point where risks exceed unity) is relatively far above the median. For instance, the saturation vapor pressure may be 100,000 times the median concentration, yet if $\sigma = 3$, the formulaic value for the mean ($\bar{x} \exp(\sigma^2/2)$) will be nearly twice as high as it “should” be if the topmost 0.1% of the PDF that lies above the saturation point were properly truncated. Conversely, although for large σ much of the mass may fall in extreme regions whose values may be highly unlikely or impossible, σ would have to be quite large indeed before one could discount the importance of less-extreme outliers. It is useful to remember that the mass of the portion of the lognormal pdf between two and three SD above the median (i.e., the region of high but probably not implausible values) increases from 14–38% of the total as σ increases from 1.0–2.5, and decreases to 14% again as σ increases from 2.5–4.0. Only for $\sigma > 4$ ($UF > 2540$) does almost all of the mass come from the extreme outliers (> 3 SD above the median).

5.4. Efficiency Advantages in Standard Computational Routines

In addition to applications where knowledge of the mass distribution is valuable *per se*, the mass distribution function can indirectly facilitate other calculations useful in risk analysis. Value-of-information (VOI) analysis is becoming a more important adjunct to risk assessment; it helps the decision-maker resolve the tension between analysis and action, and indicates which research efforts have the greatest expected return when multiple uncertainties are involved.⁽²⁸⁾ According to standard tenets of statistical decision theory, the upper bound on the value of perfect information about the “state variable” (e.g., risk, R , in attributable deaths/year) is equal to the incremental social cost one expects to incur under uncertainty, relative to the minimum cost incurred if one was clairvoyant and could choose the control strategy optimal for the true value of risk. This incremental cost will be zero in the region of the PDF for R where the control option chosen is the least costly one available (measured, perhaps, as the sum of the economic costs of risk reduction plus the health costs of the residual risks that remain). In other regions of $f(R)$, the total social cost of some other available strategy will be lower—if the true value of R is lower than expected, a strategy with lower economic costs (and lower risk reduction efficiency) would be preferable, and conversely if R is larger than expected.

If, as seems logical, the health costs are proportional to the residual risk, calculating VOI thus requires evaluating various integrals of the form $\int_a^b R f(R) dR$ and summing the incremental health costs, weighted by the probability that R will take on a value outside the region where the chosen strategy is optimal. If the uncertainty in R is lognormal, Eq. (7) allows one to substitute $E(R)[\Phi(N_a - \sigma) - \Phi(N_b - \sigma)]$ for each of these integrals. In a microcomputer environment, this expression can be evaluated via two calls to a single-valued polynomial approximation to $\Phi(\cdot)$, rather than resorting to laborious numerical integration techniques. Of course, knowledge of Eq. (7) enables one to derive a “first-cut” approximation to VOI with a hand calculator or a table of the unit normal distribution.

6. CONCLUSIONS

Exploration of the “mass density” of an uncertain and/or variable characteristic can enrich the risk management process along several dimensions. Examining

how portions of a PDF contribute to the total mass (or to the mean value) requires more conceptual sophistication than simply making a "best" or upper-bound estimate of the mass or the mean. However, to the extent that simple methods have not been available to perform such calculations, risk analysts have lacked the impetus to consider the joint effects of probability and magnitude. This paper overcomes the computational hurdle for one of the uncertainty distributions most often used in probabilistic and health risk analysis. Perhaps the ease of exposition when the PDF is lognormal will also encourage investigations into the ramifications of the mass distribution when the uncertainties in risk follow other mathematical or empirical distributions.⁽²⁹⁾

There is a growing appreciation that in our field, "a decision made without taking uncertainty into account is barely worth calling a decision."⁽³⁰⁾ Measures of "mass density" and the quantities that depend on it, such as the "conservatism" of upper-bound estimates with respect to the mean, should give more information to decision-makers who are trying to make uncertainty into an ally rather than an adversary.

ACKNOWLEDGMENTS

The helpful suggestions of Daniel Byrd, Maureen Cropper, Hadi Dowlatabadi, Michael Gough, Danny Lee, Paul Portney, Emily Silverman, and two anonymous referees are gratefully acknowledged. In particular, the advice of P. Barry Ryan was of great value during the author's doctoral studies and in the writing of this paper.

REFERENCES

1. S. Kotz, N.L. Johnson, and C.B. Read, *Encyclopedia of Statistical Sciences*, Vol. 2 (John Wiley & Sons, New York, 1982).
2. J. Aitchison and J.A.C. Brown, *The Lognormal Distribution, with Special Reference to Its Uses in Economics* (Cambridge University Press, U.K., 1957).
3. A.M. Finkel, *Uncertainty, Variability, and the Value of Information in Cancer Risk Assessment*, Sc.D. dissertation (Department of Environmental Sciences and Physiology, Harvard School of Public Health, Boston, 1987).
4. C.E. Quensel, "Studies of the Logarithmic Normal Curve," *Skandinavisk Aktuarietidskrift* **28**, 141-153 (1945).
5. A.M. Finkel, "Is Risk Assessment Really Too Conservative?: Revising the Revisionists," *Columbia Journal of Environmental Law* **14**, 427-467 (1989).
6. W.R. Ott, "A Physical Explanation of the Lognormality of Pollutant Concentrations," Presented at the 81st Annual Meeting of APCA, Dallas, Texas, June 19-24 (1988).
7. A.M. Finkel, "Estimating the Extent of Interindividual Variability in Susceptibility to Carcinogenesis: A Heterogeneity-Dynamics Approach," *Risk Analysis*, submitted (1990).
8. D.L. Freeman, R.T. Egami, N.F. Robinson, and J.G. Watson, "A Method for Propagating Measurement Uncertainties Through Dispersion Models," *JAPCA* **36**, 246-253 (1986).
9. T.E. McKone and P.B. Ryan, "Human Exposures to Chemicals Through Food Chains; an Uncertainty Analysis," *Environmental Science and Technology* **23**, 1154-1163 (1989).
10. C. Portier and D. Hoel, "Low-dose-rate Extrapolation Using the Multistage Model," *Biometrics* **39**, 897-906 (1983).
11. A.M. Finkel, "Computing Uncertainty in Carcinogenic Potency: A "Bootstrap" Approach Incorporating Bayesian Prior Information," Report to EPA Office of Policy, Planning, and Evaluation, August 8 (1988).
12. R.G. Whitfield and T.S. Wallsten, "A Risk Assessment for Selected Lead-Induced Health Effects: An Example of a General Methodology," *Risk Analysis* **9**, 197-207 (1989).
13. C. Portier and N. Kaplan, "Variability of Safe Dose Estimates When Using Complicated Models of the Carcinogenic Process: A Case Study—Methylene Chloride," *Fundamental and Applied Toxicology* **13**, 533-544 (1989).
14. K.T. Bogen and R.C. Spear, "Integrating Uncertainty and Inter-individual Variability in Environmental Risk Assessment," *Risk Analysis* **7**, 427-436 (1987).
15. S.M. Rappaport, S. Selvin, and S.A. Roach, "A Strategy for Assessing Exposures with Reference to Multiple Limits," *Applied Industrial Hygiene* **3**, 310-315. (1988).
16. A.M. Finkel, "Pitfalls of Simplified Decision-Rules for Evaluating Exposures to Lognormally-Distributed Toxicants," *Applied Industrial Hygiene* submitted (1989).
17. B.L. Cohen and N. Gromicko, "Variation of Radon Levels in U.S. Homes with Various Factors," *JAPCA* **38**, 129-134 (1988).
18. J.S. Puskin and C.B. Nelson, "EPA's Perspective on Risks from Residential Radon Exposure," *JAPCA* **39**, 915-920 (1989).
19. R.C. Schwing and D.B. Kamerud, "The Distribution of Risks: Vehicle Occupant Fatalities and Time of the Week," *Risk Analysis* **8**, 127-133 (1988).
20. A. Sen, "Poverty: An Ordinal Approach to Measurement," *Econometrica* **44**, 219-231 (1976).
21. R.A. Howard, "On Making Life or Death Decisions," in R.C. Schwing and W.A. Albers, Jr. (eds.), *Societal Risk Assessment: How Safe is Safe Enough?* (Plenum Press, New York, 1980).
22. A.B. Atkinson, "On the Measurement of Poverty," *Econometrica* **55**, 749-764 (1987).
23. L.L. Lopes, "Risk and Distributional Inequality," *Journal of Experimental Psychology: Human Perception and Performance* **10**, 465-485 (1984).
24. R. Barakat, "Sums of Independent Lognormally Distributed Random Variables," *Journal of the Optical Society of America* **66**, 211-216 (1976).
25. S.C. Schwartz and Y.S. Yeh, "On the Distribution Function and Moments of Power Sums with Lognormal Components," *Bell System Technical Journal* **61**, 1441-1462 (1982).
26. E.L. Crow and K. Shimizu (eds.), *Lognormal Distributions: Theory and Applications* (Marcel Dekker, Inc., New York, 1988).
27. G. Apostolakis and S. Kaplan "Pitfalls in Risk Calculations," *Reliability Engineering* **2**, 135-145 (1981).
28. A.M. Finkel, *Confronting Uncertainty in Risk Management: A Guide for Decision-Makers*, A Resources for the Future Report, Washington, D.C., 1990.
29. M.B. Fiering, R. Wilson, E. Kleiman, and L. Zeise, "Statistical Distributions of Health Risks," *Civil Engineering Systems* **1**, 129-138 (1984).
30. R. Wilson, E. Crouch, and L. Zeise, "Uncertainty in Risk Assessment," in D.G. Hoel, R.A. Merrill, and F.P. Perera (eds.), *Risk Quantitation and Regulatory Policy* (Banbury Report 19, Cold Spring Harbor Laboratories, Cold Spring, New York, 1985).