Whereof one cannot speak:
When input distributions are unknown

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, New York 11733

Lev Ginzburg
Department of Ecology and Evolution
State University of New York
Stony Brook, New York 11794

Resit Akçakaya
Applied Biomathematics
100 North Country Road
Setauket, New York 11733

Abbreviated title:
When input distributions are unknown

Correspondence should be directed to:
Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, New York 11733
Telephone 516-751-4350
Facsimile 516-751-3435
Internet scott@ramas.com

Accepted for publication in *Risk Analysis*

Abstract

One of the major criticisms of probabilistic risk assessment is that the requisite input distributions are often not available. Several approaches to this problem have been suggested, including creating a library of standard empirically fitted distributions, employing maximum entropy criteria to synthesize distributions from a priori constraints, and even using 'default' inputs such as the triangular distribution. Since empirical information is often sparse, analysts commonly must make assumptions to select the input distributions without empirical justification. This practice diminishes the credibility of the assessment and any decisions based on it. There is no absolute necessity, however, of assuming particular shapes for input distributions in probabilistic risk assessments. It is possible to make the needed calculations using inputs specified only as *bounds* on probability distributions. We describe such bounds for a variety of circumstances where empirical information is extremely limited, and illustrate how these bounds can be used in computations to represent uncertainty about input distributions far more comprehensively than is possible with current approaches.

KEYWORDS:

input distributions, probability bounds, maximum entropy, Laplace, ceteris paribus

Introduction

''Whereof one cannot speak, thereon one must remain silent.''
— L.J.J. Wittgenstein

''It is evidently easier for the practitioner of natural science to recognize the difference between knowing and not knowing than this seems to be for the more abstract mathematician.''
— R.A. Fisher

The difficulties of developing and justifying input distributions are well known in risk analysis and have been the subject of considerable attention recently.[1,2] While there is huge literature on the subject of estimating probability distributions from empirical data, standard approaches are of limited practical effectiveness when few data exist. In situations where data are severely limited, several strategies have been suggested to meet the challenge, including employing maximum entropy criteria[3,4] to synthesize distributions from a priori constraints, focusing on extreme value distributions[7] when the tails are of interest, collectivizing efforts to gather empirically fitted distributions,[8,2] and even using 'defaults' such as the triangular[2,1] or exponential[9] distribution whenever little empirical information is available.

We would argue that use of 'default' input distributions[2,1] should be avoided whenever possible since it has more in common with wishful thinking than it does with scientific deduction. As suggested by Wittgenstein's[10] last proposition, we have to be silent about things we cannot say anything about. While empirical efforts are always to be commended and their fruits preferred to any assumption or weaker analysis without them, we suspect it will generally be true that the gathering of relevant empirical information will lag behind pressing questions in risk assessments. In practice today, analysts often have little empirical evidence to support some of the distributions they select as inputs for probabilistic risk assessments. As a result, the analyses typically require assumptions that cannot be justified by appeal to evidence. The consequences of this may be substantial since the results of probabilistic risk analyses are known to be sensitive to the choice of distributions used as inputs,[11] an effect which is undoubtedly even stronger for the tail probabilities. The final result of any analysis can only be as good as the inputs on which it is based.

Although brute-force sensitivity studies could in principle be used to assess the robustness or fragility of the results, such studies are cumbersome to organize, computationally intense, and difficult to interpret. When many variables and assumptions are at question, the combinatorial explosion effectively prohibits any comprehensive analysis in a sensitivity study. We discuss a numerical approach that allows the calculation of bounds on arithmetic combinations of probability distributions when only bounds on the input distributions are given.

As an example, if an analyst confesses total ignorance about the variates of a particular distribution except that they must be larger than *min* and smaller than *max*, statistical convention would dictate that a uniform distribution be selected to represent this fact. This convention is very old, dating back to the beginning of probability theory, indeed to Laplace himself who argued that using any other distribution would be an expression of additional information about the relative likelihood of the possible values which, by hypothesis, is lacking. The idea has come to be known as the 'principle of insufficient reason' and has been generalized to the maximum entropy criterion.[3-6] Maximum entropy criteria allow further constraint information to be

incorporated into the argument to select an appropriate distribution that expresses only the constraints and assumes ceteris paribus (all other things being equal) in other respects about the relative probabilities of possible values.

Probability bounds

The choice of a uniform or any other particular distribution does not seem to be equivalent to the original confession of ignorance. Indeed, what is really being said by such a confession is something that is considerably weaker. We have no reason to believe that other things are equal, especially that probabilities are equal. Instead, all that is asserted is that the probabilities are unknown (ceteris incognitis). In fact, there are infinitely many distributions that are bounded by *min* and *max*. The uniform is just one of these, for which there is neither evidence nor argument. But there is something that can be said about the unknown distribution. In particular, its CDF (cumulative distribution function) must lie within a box bounded by zero and one vertically and by *min* and *max* horizontally. Such a box is depicted in Figure 1. The uniform distribution would be a diagonal line from (*min*,0) to (*max*,1) which is shown as a dotted line. Although Laplace's argument and the maximum entropy criterion would select the dotted line as the appropriate input distribution, we argue that this selection assumes more information than is actually available. All that can really be said is that the true distribution, whatever it may be, must lie within this box. That is, its CDF when plotted will lie somewhere in the area circumscribed by this box.

In more general terms, we say there exist monotonic functions *d* and *u* on [0,1] which are bounds on the inverse of the unknown distribution function *F*

$$d(p) \geq F^{-1}(p) \geq u(p)$$

where *p* is probability level. In the case of the box in Figure 1, the *u* function's graph follows the left side and top of the box. The *d* function's graph follows the bottom and right side of the box. We show below that these bounds, and refinements of them which we will presently derive, can be used in risk analyses in a way that is honest about what is known and what is unknown. Let us consider some situations that commonly arise in which only bounds on probability instead of particular distributions can be posited.

Parametric models

It is simple to compute probability bounds for many cases in which the distribution family is specified but only interval estimates can be given for the parameters. For instance, suppose that, from previous knowledge, we are willing to assume that a distribution is lognormal, but we cannot be certain about the precise values of the parameters that would define this distribution. If we have bounds on μ and σ (mean and standard deviation), bounds on the distribution can be obtained by computing the envelope of all lognormal distributions that have parameters within the specified intervals. The bounds are

$$d(p) = \max_{\alpha} L_{\alpha}^{-1}(p)$$

$$u(p) = \min_{\alpha} L_{\alpha}^{-1}(p)$$

where

$$\alpha \in \{(\mu, \sigma) \mid \mu \in [\mu_1, \mu_2], \sigma \in [\sigma_1, \sigma_2]\}$$

and *L* is the CDF of a lognormal distribution with such parameters. In principle, making these calculations might be a difficult task since α indexes an infinite set of distributions. However, in practice, finding the bounds requires computing the envelope over only four distributions: those corresponding to the parameter sets (μ_1, σ_1), (μ_1, σ_2), (μ_2, σ_1), and (μ_2, σ_2). This simplicity is the result of how the family of distributions happens to be parameterized by μ and σ. Nevertheless, it is just as easy to find probability bounds for cases with other commonly used distribution families such as normal, uniform, exponential, Cauchy, and many others.

We can also apply this approach in cases where empirical information is available. Grosof[12] suggested that standard confidence interval procedures can be used to deduce probability bounds. For instance, instead of selecting the lognormal distribution whose parameters are the best estimates from a limited empirical study, one can incorporate some of the sampling uncertainty from the study by using bounds computed from confidence intervals around the parameters. As an example, suppose that strong arguments or convincing evidence implies that a distribution is lognormal in form, with its μ and σ known only within interval ranges. Figure 2 illustrates probability bounds for the case {*shape*=lognormal, μ=[0.5, 0.6], σ=[0.05,0.1]}, for which we know the true distribution is lognormal with μ somewhere in the interval [0.5, 0.6] and σ in [0.05, 0.1]. The dotted line is the distribution that corresponds to assuming that the midpoints of the intervals are precisely the true μ and σ for the distribution.

Burmaster and Hull[13] describe a comparable way to visualize parametric uncertainty in the lognormal distribution. Their approach mixes a frequentist interpretation of probability for the underlying variable and a subjectivist interpretation for the parameters of the distribution, so that, for instance, the parameters of the lognormal distribution themselves have normal distributions. We show here that it is not essential to postulate such second-order statistical distributions where choice of distribution shape and parameters appear again in a recursive expansion of the problem. One can use the simpler construct of probability bounds to visualize the co-occurrence of variable stochasticity with parametric uncertainty.

Nonparametric models

One can also derive bounded probability regions for a variety of circumstances in which distribution shape cannot be reliably determined and empirical information is very limited. We have already discussed the case in which the only available information is {*min, max*}, for which a box generalizes the uniform distribution suggested by a maximum entropy criterion. Further information, such as knowing that the mean is a particular value, can be used to constrain this box to a smaller region. Figure 3 depicts bounds when only {*min*, *max*, *mean*} are known. The derivation of these bounds is straightforward. Let us first consider the range of *x*-values between *min* and *mean*. The upper bound over this range can be found by determining the largest possible values attained by a CDF under the specified constraints. Consider an arbitrary value $x \in$ [*min*, *mean*]. The value *p* of a CDF at *x* represents probability mass at and to the left of *x*. However much mass there is, it must be balanced by mass on the right of the mean. The greatest possible

mass would be balanced by assuming that the rest of the probability (1–*p*) is concentrated at *max*. Likewise, the arrangement of mass on the left side requires the least balance when it is all concentrated at *x*. These considerations lead to the expression

$$px + (1-p)max = mean$$

which can be solved to yield

$$p = \frac{max - mean}{max - x}$$

which specifies the largest value of the CDF for the value *x*. If there were any more probability mass at values ≤*x*, the constraint of the mean could not be satisfied by any arrangement of mass at values ≤*max*. Clearly then, the spike distributions defined by this expression describe the bounding CDF over the range [*min*, *mean*], subject to the fundmental constraint 0≤*p*≤1. The position of the lower bound is determined by the degenerate distribution which has all its mass at the mean. The CDF for this distribution follows the *x*-axis from *min* to *mean*. Lower and upper bounds for values larger than the mean can be derived by similar (but upside-down) arguments. The resulting bounds are optimal in the sense that they could not be any tighter without excluding at least some portion of a CDF from a distribution satisfying the specified constraints. It is important to understand, however, that this does not mean that any distribution whose CDF is inscribed within this bounded probability region would satisfy the constraints.

The most potent information is knowledge of medians which pinches the uncertainty to a definite point. When information is limited to {*min*, *max*, *median*}, we obtain bounds on probability such as are depicted in Figure 4. Having reliable knowledge of other percentiles would correspond to similar points at other probability levels through which we can be sure the true distribution, whatever it is, must pass.

If one can assume that the underlying distribution is unimodal and that reliable estimates are available for {*min*, *max*, *mode*}, then the probability bounds are depicted in Figure 5. Again, we emphasize that not every curve contained in this region satisfies the given constraints. However, the bounds are optimal in the sense that they could not be tighter without excluding some distribution that does satisfy the specified constraints.

As with parametric models, it will usually be the case that the mean, percentiles, and other parameters are only imperfectly known. This suggests that confidence intervals rather than point estimates should be used for them. Likewise, when *min* and *max* are determined by empirical observation rather than by mathematical argument or theoretical consideration, they are better estimated as confidence intervals whenever point estimates would imply a false certainty. The envelope of all possible underlying distributions can be calculated as the union of bounded probability regions in a manner analogous to that used above to compute probability bounds for parametric models. Again, in many cases, the calculation of the bounds is a fairly simple matter even though it formally implies consideration of an infinite family of curves. All of the cases derived so far, for instance, are easy to generalize so that any of the numeric values within the braces may be given either as a fixed number or an interval.

Further cases may also be derived for simultaneous multiple constraints by intersecting the bounded probability regions obtained for each constraint separately. The bounds would be

$$d(p) = \min_i d_i(p)$$

$$u(p) = \max_i u_i(p)$$

where *i* indexes the constraints. For instance, if one knows the *min* and *max* values and has reliable estimates of both the mean and the mode, it is justifiable to construct probability bounds for this case by intersecting the regions described for {*min, max, mode*} and {*min, max, mean*}. Although this approach will often yield optimal bounds, it is not guaranteed to do so. It cannot take into account constraints that are not only simultaneous, but interacting. For instance, if we know only {*min, max, mean=median*} or {*min, max, median=mode*}, intersecting the bounded regions will not in general yield optimal bounds, which must be derived separately for these cases.

Empirical models

We can also use Kolmogorov-Smirnov confidence limits[14] on an empirical histogram to construct input distributions. This method requires the analyst to specify both the support (outside of which the distribution is truncated) and the confidence level to be used. Figure 6 illustrates an empirical histogram composed from a sample of six points (.2, .5, .6, .7, .75, and .8) for a variable constrained between zero and one. The naive histogram is shown as a dotted line and the 95% Kolmogorov-Smirnov limits are shown as solid lines. The bounds would be tighter if there were more samples or we used a lower confidence level. We note that exactly how the dotted line in Figure 6 should be drawn is the subject of some controversy. Recent research using Monte Carlo simulation shows that the Kolmogorov-Smirnov bounds may be wider than necessary. A maximum entropy solution for a given set of empirical data is discussed by Solana and Lind.[15]

Assumed distributions

Of course, an analyst may be legitimately confident about the shapes and parameters of some input distributions on account of good empirical evidence. This confidence can be represented by using osculating bounds

$$d(p) = F^{-1}(p) = u(p)$$

describing the distribution function. We have implemented in software most of the commonly used distributions, including Bernoulli, Cauchy, Dirac's delta, discrete uniform, exponential, geometric, Gumbel, Laplace, logistic, lognormal, logtriangular, loguniform, normal, Pareto, power function, Rayleigh, triangular, uniform, and Weibull. For distributions with theoretically infinite supports, the software permits automatic truncation at the 100α% and $100(1-\alpha)$% percentiles where α can be selected by the analyst.

Computing with probability bounds

The input distributions used in a probabilistic risk assessment need *not* be particular, well-defined statistical distributions when there is insufficient empirical information available to justify their selection. Williamson and Downs[16] gave numerical methods for computing bounds on the result of addition, subtraction, multiplication and division of random variables when only bounds on the input distributions are given. Suppose that variables A and B have bounds (d_A,u_A) and (d_B,u_B) respectively, and that each of these four functions is evenly discretized into m+1 elements. Assuming A and B are independent, the bounds on the sum $A+B$ have a discretization

$$d(i/m), \quad u(i/m) \qquad i \in [0,m]$$

where $d(i/m)$ is approximated by the $(i+im+m)$th element of a numerical sorting of the $(m+1)^2$ values

$$d_A(j/m) + d_B(k/m) \qquad \forall j,k \in [0,m]$$

and $u(i/m)$ is approximated by the $(i+im)$th element of a numerical sorting of the values

$$u_A(j/m) + u_B(k/m) \qquad \forall j,k \in [0,m].$$

The algorithm for subtraction is virtually the same except that the pluses between the d's and the u's are replaced by minuses. Multiplication and division use their respective operators too, so long as both variables are strictly positive. A more elaborate algorithm is required in the general case, although division is undefined whenever the support of the divisor includes zero. Aside from the four basic arithmetic operations, it is also possible to handle other functions as well, including minimum, maximum, logarithms, exponentiation and integral powers.

Williamson and Downs[16] also describe numerical methods for computing bounds *without* using an assumption of independence between the variables. Bounds on the sum of A and B, for example, are

$$d(i/m) = \min_{j=i}^{m} (d_A(j/m) + d_B((i-j+m)/m))$$

$$u(i/m) = \max_{j=0}^{i} (u_A(j/m) + u_B((i-j)/m))$$

where i varies between 0 and m. These bounds are guaranteed to enclose the true answer no matter what correlation or statistical dependency exists between A and B. Similar expressions can be used for subtraction, multiplication, division and other mathematical operations. These methods constitute a comprehensive dependency bounds analysis.[17]

These algorithms are very general and can be used for practically any distribution that has a finite support (which means that, as for Monte Carlo methods, infinite distributions must be truncated to a finite range). The resulting bounds can also be used in subsequent calculations. For instance, the bounds for $A+B$ may be combined with C where C is a particular CDF or bounds on a CDF. This numerical method can therefore be used in risk analysis to compute bounds on

probability distributions.[17-19] If the mathematical expression contains only one instance of each variable, then the resulting bounds are furthermore the best bounds possible.[20] In any case, the bounds on the probabilities are guaranteed to be conservative. This is to say that, if the model is correct and the input bounds enclose the true distributions of the inputs, then the bounds obtained by this approach are guaranteed to enclose the true answer.

Since the bounds computed from these arithmetic operations are guaranteed by construction to enclose the true answer, we can analyze the bounds to discover facts about the distribution of the true result, even without computing it directly. These facts are usually also in the form of constraints and inequalities. For instance, from bounds on the distribution function, one can compute bounds on the distribution's mean. One can also compute bounds on the variance of the sum, as well as higher-order functions, although these are often too wide to be of practical interest. Often the most interesting inequalities pertain to the bounds on the tails of the distribution.

Distribution tails, which are often the focus of regulatory concern,[7] can be and often are very sensitive to the shapes of input distributions. This method gives analysts a computational way to avoid underestimating tails even when they cannot supply much empirical information about the input variables. Although the results are conservative with respect to the professed ignorance about input distributions, they do not appear to be hyperconservative[21] to the degree observed in many worst cases analyses. This method yields final results whose supports are identical in breadth to those that would be obtained by a Monte Carlo analysis. While they can strongly disagree about the magnitude of the tail probabilities, truly impossible outcomes are excluded by both methods. If, as Cullen[22] suggests, there are commonly only a handful of variables whose uncertainties are of sufficient breadth to strongly influence the dispersion of the answer, then it is perhaps even less likely that this approach could compound conservativisms into hyperconservativism.

A numerical example

To illustrate the use of the probability bounds in risk assessments, we compute the bounds on an arithmetic expression involving random variables whose distributions are incompletely specified. For instance, consider the sum of some of the distributions whose bounds are illustrated in this paper

$$A + B + C + D$$

where

A = {*shape*=lognormal, μ=[0.5, 0.6], σ=[0.05, 0.1]},

B = {*min*=0, *max*=0.4, *mode*=0.3},

C = {*min*=0, *max*=1, *data*=(0.2, 0.5, 0.6, 0.7, 0.75, 0.8)}, and

D = Uniform(0,1).

Thus, we will 'add' four pairs of probability bounds constructed in disparate ways. In this

calculation we will assume the true underlying distributions for these four variables are independent (although we could have omitted this assumption). The result is displayed in Figure 7 as a pair of solid curves which bound the distribution of this sum. For comparison, the result that would be obtained from a calculation using specific input distributions selected under maximum entropy criteria is also displayed as a dotted line. The mean of the true distribution of the sum must lie somewhere in the interval [1.4, 2.3]. The most important information from a risk analyst's perspective is likely to be the bounds on the tail probabilities. Whatever the true inputs are, the distribution of the result has a minimum that is no smaller than about 0.5, and a maximum that is no larger than about 3. One can see very clearly from this small example that fully accounting for the uncertainty in input distributions yields a very different picture with respect to the tails of the distribution. In particular, the probability of extreme values is seen to potentially much greater than was predicted by the maximum entropy solution, although that we cannot generally conclude that they will be. The uncertainty in these probabilities is a direct consequence of the uncertainty in the input distributions.

Although calculating with probability bounds yields a very different picture of the tail probabilities, this numerical experiment also illustrates that even rather loose bounds do not radically increase the uncertainty in the final result. In fact, there is surprisingly little uncertainty in the final result given that we have used these probability bounds for inputs rather than the maximum entropy inputs. Nevertheless, it seems possible that an analyst's decisions about the supports of the input distributions (i.e., where the *max* and *min* are) could influence the behavior of the result's tails. The sensitivity of the tails to such decisions is unknown, and further research is required to understand the relationship.

Relationship with interval analysis

We note that the method still works even if we have no information about the input distributions except their supports. Even if we only know the minimum and maximum possible values, and have no further information that would permit us to draw any conclusions whatever about the probabilities of the values, we can still compute bounds on expressions involving sums, products, differences, quotients, logs, powers, etc. In fact, calculation with these 'probability boxes' is exactly equivalent to interval analysis.[23,24] This finding is in line with other observations by Williamson[25] that show the connection between probabilistic arithmetic and interval analysis. Interval analysis (which includes any well-formulated approach based on worst case analysis) is seen to be a special case of probability theory. Although they have often been thought too disparate in focus and methodology to be combined in any meaningful way, this explains how to integrate worst case and probabilistic risk analyses.

Limitations of using probability bounds

Technical limitations. There are three potential difficulties with the approach we suggest. First, the algorithms given here do not always work for multiplication and division when distributions have ranges that include non-positive numbers. Second, although the algorithms can either assume independence or make no assumption about dependency, they cannot make use of information about non-zero correlations among the variables. Third, optimally narrow solutions are no longer guaranteed when repeated variables appear in the risk expressions. These limitations may be surmountable, but will require further methodological research.

Need to derive bounds for new knowledge sets. We have described probability bounds for a variety of common situations, including most of those satisfied by maximum entropy criteria. However, several other situations are commonly encountered for which we have not given bounds. Prominent among these are situations for which only measures of central tendency and dispersion are known. For instance, what bounds are appropriate when one only knows {*mean*, *variance*}? We defer the derivation of these and related bounds to a future publication.

The simple ways to derive bounds for new situations such as intersection and union of bounding regions do not provide a complete apparatus for all circumstances. It may be difficult to take into account other kinds of information or assumptions that analysts are willing to make. For instance, how does one derive optimal bounds when we know the mean and variance and that distribution is symmetric? What are the best bounds when one knows the mean and that the mode is less than the mean?

Non-graded treatment of uncertainty. Uncertainty about the input probability distributions is treated as a kind of interval uncertainty. This means that one makes no statement about exactly where within the bounds the true distribution lies, nor even in which areas it is more likely to be. The approach cannot handle second-order probability distributions[26-28,13] (where the parameters of a distribution are themselves drawn from probability distributions). Analysts who believe that uncertainty and variability need to be described in this way, and also believe that they can accurately specify these second-order distributions, will not be able to use the proposed method to take advantage of this information, except perhaps in nested, multiple analyses that will likely grow as complicated as sensitivity studies.

Incomplete treatment of model uncertainty. Although these methods can be used to treat model uncertainty that is reflected in the choice of input distributions, there is usually a large measure of further uncertainty about the model that cannot be incorporated into this approach, including indecision about the mathematical form of the model, choices about which functions to use and the appropriate level of modeling abstraction. Comprehensively exploring the consequences of such uncertainty will still require a sensitivity study.

Results limited to probability bounds. The chief limitation of this approach may be that it supplies no specific answer, but only bounds on the answer. Some would argue that this is because it is in fact impossible to compute a precisely characterized probability distribution in the face of real ignorance of the kind encountered in risk analysis. Others, however, do not hesitate to employ maximum entropy criteria or other rules of thumb to arrive at such distributions. We expect that even analysts who feel it important to compute a precise distribution (or even a single scalar answer) will nevertheless appreciate this approach as a useful tool for conducting quality assurance checks on their analyses.

Discussion: ceteris paribus versus ceteris incognitis

The idea that it is appropriate to assume probabilistic uniformity in circumstances with no evidence to the contrary dates to the beginning of probability theory. Of course, we are by no means the first to question its appropriateness. It was rejected outright by George Boole[29] in his *Laws of Thought*. As Ronald Fisher[30] recounts, its wide acceptance has mostly been the result of simply not knowing what else to do. The need to develop prior probabilities for Bayesian methods *before* data were collected has perhaps also contributed to a general confusing of 'ceteris

paribus' with 'ceteris incognitis'. Computing with probability bounds allows us now to make progress in risk calculations without making this certainly questionable—and probably often false—assumption.

To be fair, we should note that the proponents of maximum entropy criteria are not pretending that the selected distribution represents variability that actually exists.[31] They only assert that this distribution is the optimal way to represent the subjective ignorance about the underlying quantity. We do not contest that it is the optimal way to represent ignorance in a single probability distribution, but we argue that it is not a comprehensive way to express this ignorance. Now that methods for calculating with bounds on probability distributions are available, there is no necessity of selecting a single distribution for each input.

The alternative methods for comprehensively handling uncertainty about input distributions in probabilistic risk analyses involve complicated sensitivity studies or 'two-dimensional' probabilistic risk assessments.[26-28] Both of these methods are often cumbersome and can be difficult to interpret. Moreover, they confound frequentist and subjectivist theories of probability. Probability bounds allow a full range of arithmetic calculations under a purely classical freqentist interpretation of probability.[32]

The approach we have suggested in this paper allows analysts to make probabilistic risk assessments even when extremely little reliable empirical information is available about the input distributions. The approach will also be useful when empirical information is abundant because it allows sampling uncertainty expressed as confidence limits on probability distributions or the parameters that define them to be used in calculations. Indeed, different variables in an analysis can be given either as particular well-specified distributions or as bounds on distributions as appropriate to represent the state of *empirical information* available about each variable. The degree of specificity does not affect how the variables are combined in subsequent arithmetic operations, and the result thereby represents the appropriate level uncertainty rather than a false precision gained by unjustified assumptions. This approach therefore provides a way to honestly and fully characterize the uncertainty in the input distributions used in risk analyses which, as has argued on scientific[33,34] as well as ethical[35] grounds, is an important obligation of the risk analyst.

Acknowledgements

This paper benefitted from discussions with Robert C. Lee (Golder Associates), Clark Carrington (U.S. Food and Drug Administration), David Burmaster (Alceon Corporation), Daniel Wartenberg (Rutgers) and Thomas F. Long (Illinois Department of Public Health and ChemRisk). This publication was made possible in part by funding from a Small Business Innovation Research grant (1R43ES06857) to Applied Biomathematics from the National Institute of Environmental Health Sciences (NIEHS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily reflect official views of NIEHS or NIH.

References

1. Y.Y Haimes, T. Barry and J.H. Lambert, ''When and how can you specify a probability distribution when you don't know much?'' *Risk Analysis* **14**, 661-706 (1994).

2. B. Finley, D. Proctor, P. Scott, N. Harrington, D. Pasutenbach and P. Price ''Recommended distributions for exposure factors frequently used in health risk assessment,'' *Risk Analysis* **14**, 533-553 (1994).
3. R.C. Lee and W.E. Wright, ''Development of human exposure-factor distributions using maximum-entropy inference,'' *Journal of Exposure Analysis and Environmental Epidemiology* **4**, 329-341 (1994).
4. E.T. Jaynes, ''Information theory and statistical mechanics,'' *Physical Review* **106**, 620-630 (1957).
5. R.D. Levine and M. Tribus, *The Maximum Entropy Formalism*. (MIT Press, Cambridge, 1978).
6. W.T. Grandy, Jr. and L.H. Schick, *Maximum Entropy and Bayesian Methods*. (Kluwer Academic Publishers, Dordrecht, 1991).
7. J.H. Lambert, N.C. Matalas, C.W. Ling, Y.Y. Haimes and D. Li ''Selection of probability distributions in characterizing risk of extreme events,'' *Risk Analysis* **14**, 731-742 (1994).
8. A.C. Cullen and H.C. Frey, *Developing Distributions for Use in Probabilistic Exposure Assessments*. [To appear] (1995).
9. P. Bratley, B.L. Fox and L.E. Schrage, *Guide to Simulations*. (Springer-Verlag, New York, 1983).
10. L.J.J Wittgenstein, ''Tractatus logico-philosophicus,'' *Annalen der Naturphilosophie* (1921).
11. J. Bukowski, L. Korn and D. Wartenberg, ''Correlated inputs in quantitative risk assessment: the effects of distributional shape.'' *Risk Analysis* [in press].
12. B.N. Grosof, ''An inequality paradigm for probabilistic knowledge: the logic of conditional probability intervals,''. *Uncertainty in Artificial Intelligence*. L.N. Kanal and J.F. Lemmer (eds.), Elsevier Science Publishers, Amsterdam (1986).
13. D.E. Burmaster and D.A. Hull, ''Lognormal probability plots as a way to distinguish and visualize variability and uncertainty in a lognormal random variable,'' [To appear] (1995)
14. R.R. Sokal, and F.J. Rohlf *Biometry*. (Freeman and Company, San Francisco, 1981).
15. V. Solana and N.C. Lind, ''Two principles for data based on probabilistic system analysis,'' *Proceedings of ICOSSAR '89, 5th International Conferences on Structural Safety and Reliability*. (American Society of Civil Engineers, New York, 1990).
16. R.C. Williamson and T. Downs, ''Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds,'' *International Journal of Approximate Reasoning* **4,** 89-158 (1990).
17. S. Ferson and T.F. Long, ''Conservative uncertainty propagation in environmental risk assessments,'' *Environmental Toxicology and Risk Assessment - Third Volume, ASTM STP 1218*, J.S. Hughes, G.R. Biddinger, and E. Mones (eds.) (American Society for Testing and Materials, Philadelphia, 1994).
18. S. Ferson and M. Burgman, ''Correlations, dependency bounds and extinction risks,'' *Biological Conservation* [in press] (1995).
19. S. Ferson, ''Naive Monte Carlo methods yield dangerous underestimates of tail probabilities,'' *Proceedings of the High Consequence Safety Symposium, Sandia National Laboratories* [in press] (1995).
20. R.C. Williamson, *Probabilistic Arithmetic*. (Ph.D. dissertation, University of Queensland, 1989).
21. D.E. Burmaster and R.H. Harris, ''The magnitude of compounding conservatisms in superfund risk assessments,'' *Risk Analysis* **13,** 131-134 (1993).
22. A.C. Cullen, ''Measures of conservatism and probabilistic risk assessment.'' Society for Risk Analysis Annual Meeting, Savannah, Georgia (1993).
23. R.E. Moore, *Interval Analysis* (Prentice-Hall, Englewood Cliffs, New Jersey, 1966).

24. A. Neumaier, *Interval Methods for Systems of Equations*. (Cambridge University Press, Cambridge, 1990).
25. R.C. Williamson, ''An extreme limit theorem for dependency bounds of normalized sums of random variables,'' *Information Sciences* **56**, 113-141 (1991).
26. C.D. Carrington and P.M. Bolger, ''Two-dimensional Monte-Carlo simulations: uses and technique. Society for Risk Analysis Annual Meeting, Final Program and Abstracts, D-POSTER-05, (1993).
27. J.C. Helton, ''Treatment of uncertainty in performance assessments for complex systems,'' *Risk Analysis* **14**, 483-511.
28. F.O. Hoffman and J.S. Hammonds, ''Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability,'' *Risk Analysis* **14**, 707-712 (1994).
29. G. Boole, G. *An Investigation of the Laws of Thought, On Which Are Founded the Mathematical Theories of Logic and Probability*. (Walton and Maberly, London, 1854).
30. R.A. Fisher, *Statistical Methods of Scientific Inference*. (Hafner Press, New York, 1973).
31. S.F. Gull, ''Some misconceptions about entropy,'' in B. Buck and V.A. Macauley (eds.), *Maximum Entropy in Action*. (Oxford Science Publications, Oxford, 1991).
32. P. Walley, and T.L. Fine, ''Towards a frequentist theory of upper and lower probability,'' *Annals of Statistics* **10**, 741-761 (1982).
33. A. Finkel, *Confronting Uncertainty in Risk Management*. (Resources for the Future, Washington, 1990).
34. G.M. Gray, ''The challenge of risk characterization,'' Report to the Office of Technology Assessment, U.S. Congress. National Technical Information Service Publication PB93-218857 (1993).
35. K. Shrader-Frechette, *Ethics of Scientific Research*. (Rowman & Littlefield, Lanham, Maryland, 1994).

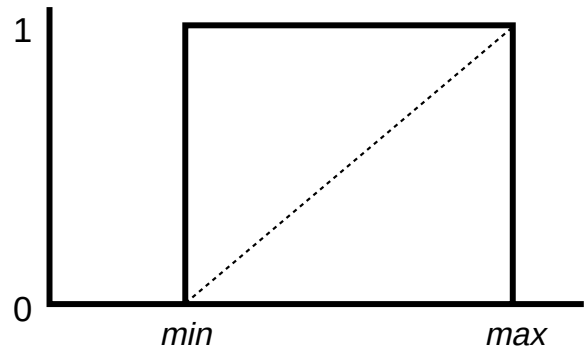

Figure 1. Bounds (solid lines) on the cumulative distribution function (CDF) of any probability distribution whose support ranges between *min* and *max*. The dotted line is the CDF of the uniform distribution which is the one chosen by a maximum entropy criterion or the principle of insufficient reason from among the infinite number of distributions bounded in this box.

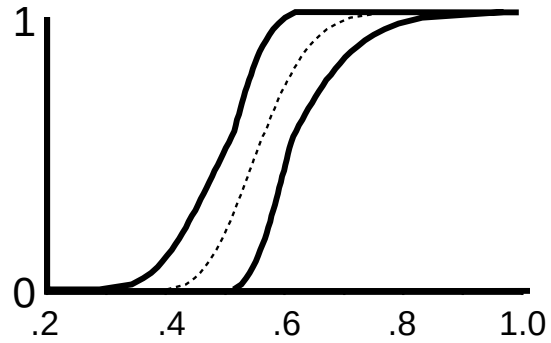

Figure 2. Bounds on the CDF of a lognormal distribution having μ somewhere in the interval [0.5, 0.6] and σ somewhere in [0.05, 0.1]. The dotted line is the CDF for the one with μ=0.55 and σ=0.075.

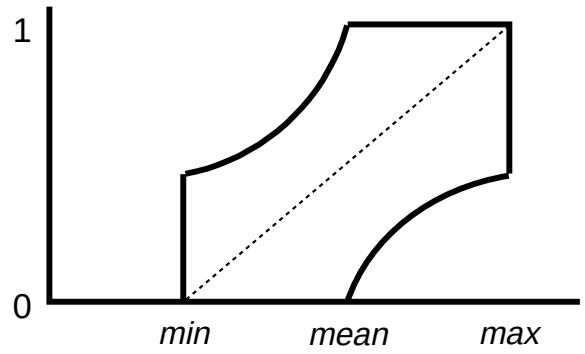

Figure 3. Bounds on the CDF of a distribution with support on [*min*, *max*] and a central tendency of *mean*. The upper and lower bounds are given by $\langle (max-mean)/(max-x) \rangle$ and $\langle (x-mean)/(x-min) \rangle$ respectively, where the angle brackets constrain the value to the interval [0,1]. The dotted line denotes the maximum entropy solution which in general is a beta distribution.

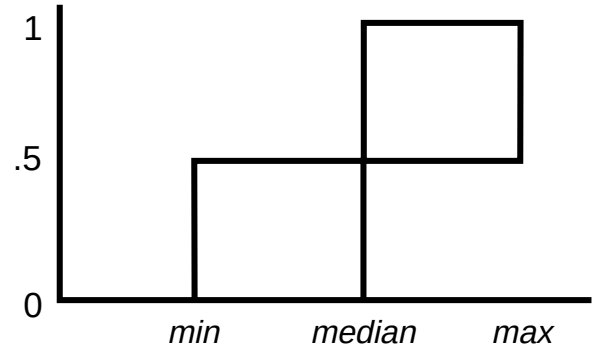

Figure 4. Bounds on the CDF of a distribution with the indicated support and median.

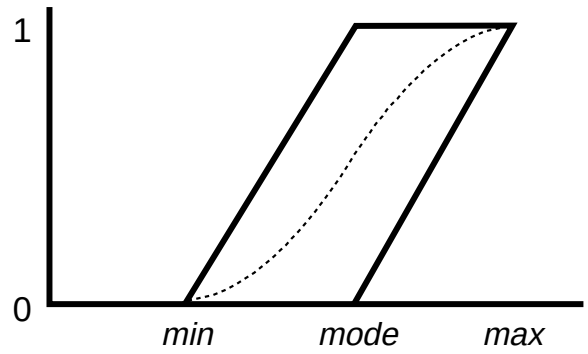

Figure 5. Bounds on the CDF of a distribution with the indicated support and mode. The bounds are given by $\langle (x-min)/(mode-min) \rangle$ and $\langle (x-mode)/(max-mode) \rangle$, limited to values on [0,1]. The dotted curve denotes the CDF of a triangular distribution which is often used when only the minimum, maximum and mode of a distribution are known.

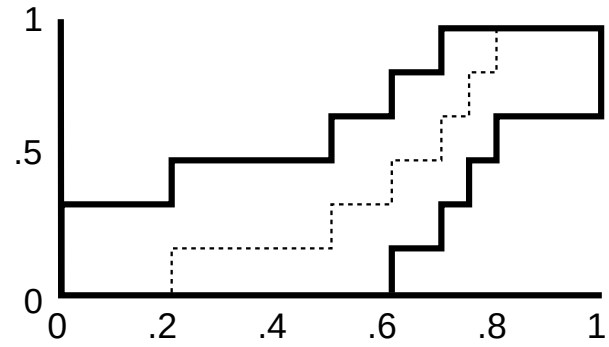

Figure 6. Kolmogorov-Smirnov 95% confidence bounds (solid lines) on an empirical distribution (dotted line) made up of 8 values (0.2, 0.5, 0.6, 0.7, 0.75, and 0.8) assuming the distribution's support is [0,1].

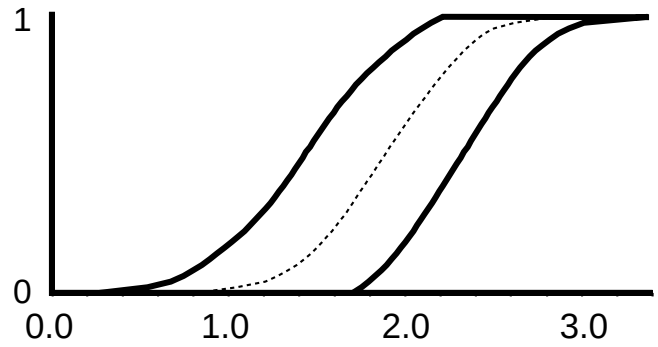

Figure 7. Bounds on the distribution of the sum of four random numbers whose distributions are described by bounds. The dotted curve is the distribution that would be obtained had the uncertainty in the probability distributions been ignored. See text.