Improving Analytical Understanding Through the Addition of Information: Bayesian and Hybrid Mathematics Approaches

J. Arlin Cooper
Sandia National Laboratories
Albuquerque, NM 87185-0490

Kathleen V. Diegert
Sandia National Laboratories
Albuquerque, NM 87185-0829

Abstract

Safety analysts frequently must provide results that are based on sparse (or even no) data. When data (or more data) become available, it is important to utilize the new information optimally in improving the analysis results. Two methods for accomplishing this purpose are Bayesian analysis, where "prior" probability distributions are modified to become "posterior" distributions based on the new data, and hybrid (possibilistic/probabilistic analysis) where possibilistic "membership" portrays the subjectivity involved and the probabilistic analysis is "frequentist." Each of these approaches has interesting features, and it is advantageous to compare and contrast the two. In addition to describing and contrasting these two approaches, we will discuss how features of each can be combined to give new advantages neither offers by itself.

1 Introduction

It is not unusual for safety analysts to provide results that are based on sparse (or even no) definitive data. When the pertinent data start becoming available, it is important to utilize the new information optimally. The obvious goal is to provide the best possible analysis results.

* Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Two methods for accomplishing this purpose are Bayesian analysis, where "prior" probability distributions are modified to become "posterior" distributions based on the new data, and hybrid (possibilistic/probabilistic analysis) where possibilistic "membership" portrays the subjectivity involved, the probabilistic analysis is "frequentist," and the amount of subjectivity involved is indicated. Each of these approaches has interesting features, and it is advantageous to compare and contrast the two. Although not directly part of this study, we also considered Dempster-Shafer theory, from which Bayesian theory and possibility theory can be derived as special cases. However, the approaches we selected appeared to fit better with our requirement to separately indicate the relative contributions of data and judgment.

2 Bayesian Analysis

The Bayesian approach [1] is well known. Based on Bayes' theorem, posterior density is a function of the prior distribution assumed and the joint density of the new data. This approach can also be used for model uncertainty by incorporating expert belief about the validity of various models, or alternatively, model uncertainty can be treated as a sensitivity issue. A specific form of prior, called a "conjugate prior" is often assumed. For example, a Beta distribution or Gamma distribution may be appropriate "informative" priors. A uniform or loguniform may be used as a "noninformative prior." The Bayesian approach has the general attributes of rapid convergence as meaningful data are accumulated, and a relatively efficient determination of output information from input information. The disadvantages are that if new data are not obtained, or if the new data are sparse, the output accuracy is only as good as the assumed prior.

Equation 1 (Bayes Theorem) demonstrates the derivation of a posterior density function from a prior density and observed data.

$$g(x \mid d_1, d_2, \ldots, d_i) = h(x) \frac{f(d_1, d_2, \ldots, d_i \mid x)}{f(d_1, d_2, \ldots, d_i)} \tag{1}$$

where $h(x)$ is the prior density, and $f(d_1, d_2, \ldots, d_i)$ is the density function of the observed data.

As an example [1], assume we want to estimate the occurrence rate λ for a two-truck accident involving a DOE cargo transportation vehicle during a trip on Highway No. 1 (length L). We initially have data on two-truck accidents pertaining to all U.S. highways normalized to a trip of length L, have justification for a gamma prior distribution, and determine that the expected value is 5×10^{-4}, with variance 2.5×10^{-4}. We calculate an initial estimate for the gamma distribution parameters:

$$\beta = 0.001; \alpha = 2,$$

where the gamma density function is:

$$f(\lambda) = \frac{\alpha^{\beta}}{\Gamma(\beta)} \lambda^{\beta-1} e^{-\alpha\lambda}$$

The parameters α and β are determined from the fact that the mean and variance of a gamma distribution are respectively β/α and β/α^2 If we then obtain actual data for Highway 1 that there have been 21 two-truck accidents in 174,992 trips, the posterior distribution is gamma with parameters:

$$\beta = 42.001; \alpha = 174{,}994.$$

resulting in an expected value of 2.4×10^{-4}. If subsequent data on DOE transportation vehicles becomes available, the new information will again be processed by the Bayes equation.

3 Hybrid (Probabilistic/Possibilistic) Knowledge Apportionment

The hybrid approach [2] represents initial subjective belief as a possibilistic function rather than a probabilistic function. The difference from the previous approach is that possibilistic membership functions directly describe the vagueness or the imprecision in the measurement or the representation without indicating any probability of any particular value. Also included in the hybrid approach is a "scale factor" that represents the relative amount of reliance on information in each part (possibilistic and probabilistic) of the representation. As new information becomes available, the scale factors change to portray the relative amount of new knowledge, and the probabilistic function begins to converge on the most appropriate representation. The advantages of this approach are that the amount of subjectivity involved is clearly indicated and no prior distribution (which could be initially incorrect) need be assumed. The disadvantages are that multiple components of information must be assembled by the analysis recipient, and the probabilistic result does not benefit from Bayesian efficiency.

Returning to the previous example, the extent of knowledge about the problem is fractionally partitioned between stochastic and subjective portions by a hybrid number according to a scaling fraction:

$$h(x) = a \times p(x) + (1-a) \times f(x) \qquad (2)$$

where $p(x)$ is a probability distribution, $f(x)$ is a possibilistic function, a is an estimated scale factor representing the fractional stochasticity of the overall knowledge ($0 \le a \le 1$), and where $\times$ and $+$ are abscissa operators (on x values). We compute probabilistic and possibilistic portions separately and combine the two results in a hybrid representation. The use of hybrid formulations in this approach is mainly to associate the separate views in a more informative entity.

When the values of input variables are not well known, risk analysts may expect to improve their analyses by incorporating new information that is learned through additional tests, accident assessments, etc. In a Bayesian sense, stochastic information can be improved. However, there are significant differences between Bayesian and possibilistic analyses. The general effect of Bayesian and other forms of probabilistic analysis is that extremes (tails of the distributions) are suppressed relative to the results of possibilistic analysis [3]. Since new input data may only slightly improve the stochastic knowledge about ill-defined situations such as abnormal environment responses, a non-Bayesian hybrid analysis has a useful role.

For the example under consideration, our first probabilistic data are the same as above (U.S. highway data). For possibilistic data, assume we have surveyed "experts" on transportation safety, specifically including judgment on how special DOE transportation controls might reduce the accident frequency from the general state. For illustration, assume the widest uncertainty bounds of the estimated frequency are 10^{-5} to 10^{-2}, and the narrowest bounds are 5×10^{-5} to 5×10^{-3}. The possibilistic function is shown in Fig. 1. [Note that if the expert opinions were inconsistent, the more general case of a "belief" function would be required.] The estimated value of the U.S. highway data to the problem of interest is a=0.3. This means that the recipient of the hybrid information is expected to place about 70% weight on the subjective (possibilistic) inputs and 30% weight on the U.S. highway data.

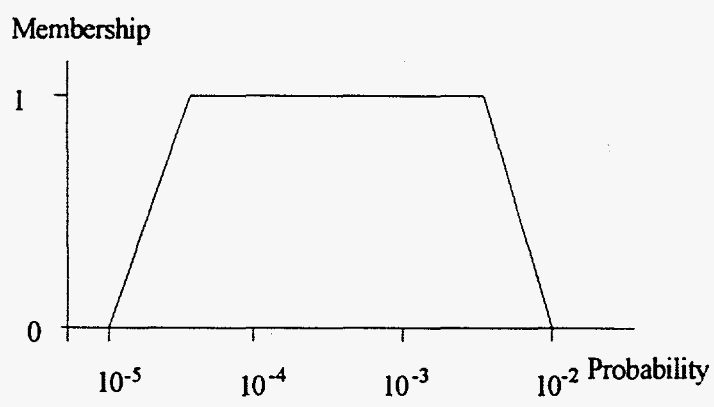

Figure 1. Possibility Function for Example

When the Highway 1 data become available, the probabilistic data are used directly:

$$\beta = 42; \alpha = 174{,}992.$$

Based on the new knowledge, expert judgments are used to obtain a new possibilistic function specific to DOE transportation vehicles (Fig. 2).

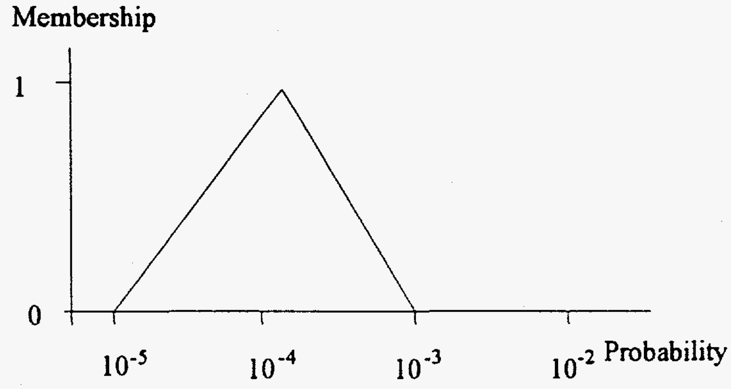

Figure 2. Updated Possibilistic Estimate for Example

The variation from the Highway 1 data (which have an extremely narrow variance and a slightly higher expected value than that corresponding to the peak value in Fig. 2) is because of uncertainty about the applicability of past Highway 1 general population data to future DOE transportation scenarios. The scale factor information is updated to 0.7. For these hybrid data, the recipient should weight the Highway 1 data about 70% and the possibilistic judgment about 30%. When actual DOE transportation data on Highway 1 become available, updates in all three hybrid parameters can be made.

4 Combining Bayesian and Hybrid Approaches

Advantages of combining the Bayesian approach with the hybrid approach are that strengths are emphasized and weaknesses are minimized. Returning to the process described in Sections 2 and 3, the major change from Section 3 implied by the combined approach is that new probabilistic data are used through Bayesian updating rather than directly. For the example, the difference between the direct Highway 1 data and the Bayesian update from U.S. highway data to incorporate Highway 1 data is insignificant. The Section 3 distribution would be:

$$\beta = 42.001; \alpha = 174{,}994.$$

However, since the combined technique is more comprehensive than the strict application of the hybrid approach, we find the combination very intriguing for future applications.

Conclusions

Since safety analysis problems frequently involve sparse data techniques for initial risk probabilities and incorporation of new data must be considered carefully. Both Bayesian estimates and incorporation of new data and hybrid incorporation of subjective information and new data are useful. Combining the two approaches may be even more useful for the safety analysis community.

References

1. Diegert, K. V., and K. M. Hansen, "Methods of Uncertainty Analysis for Probabilistic Risk Assessment," Sandia National Laboratories Draft to be Published as *DOE Safety Guide 201—Uncertainty Analysis*, 1992

2. Cooper, J. A., Scott Ferson, and Lev Ginzburg, "Hybrid Processing of Stochastic and Subjective Uncertainty Data," *Risk Analysis*, Vol. 16, No. 6, Dec. 1996

3. Cooper, J. A., "Fuzzy-Algebra Uncertainty Analysis for Abnormal-Environment Safety Assessment," *Journal of Intelligent and Fuzzy Systems*, Vol. 2, No. 4, 1994