2006

# Partial information use in uncertainty quantification

Jianzhong Zhang
*Iowa State University*

# Partial information use in uncertainty quantification

by

Jianzhong Zhang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Engineering

Program of Study Committee:
Daniel Berleant (Major Professor)
Gerald Sheblé
Soumendra Nath Lahiri
Zhengdao Wang
Yao Ma

Iowa State University

Ames, Iowa

2006

Graduate College

Iowa State University

This is to certify that the doctoral dissertation of

Jianzhong Zhang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

# TABLE OF CONTENTS

# LIST OF FIGURES

# ABSTRACT

Uncertainty exists frequently in our knowledge of the real world. Two forms of uncertainty are considered. One is variability coming from stochasticity. The other is epistemic uncertainty, also called 2nd order uncertainty and other names as well. Often it comes from ignorance or imprecision. In principle, this kind of uncertainty can be reduced by additional empirical data.

Stochasticity is well studied in the field of probability theory. A variety of methods have been developed to address epistemic uncertainty. Some of these approaches are confidence limits, discrete convolutions, probabilistic arithmetic, Monte Carlo simulation, copulas, stochastic dominance, clouds, and distribution envelope determination. Belief and plausibility curves, upper and lower previsions, left and right envelopes and probability boxes designate an important type of representation for bounded uncertainty about distribution.

Some methods combine probability theory and interval mathematics. Intervals have the potential for bounding the result of an operation. Discretization error coming from discretizing distributions may be bounded by intervals. Distribution envelope determination (DEnv) uses interval based analysis. If the dependency is not specified, result bounds will include the entire range of possible dependencies. These bounds will be wider than if a particular dependency is specified. I have worked on new algorithms to process the dependency relationships. Pearson correlation can be used to improve the results, for example. Also partial dependence information might be available in the form of unimodality or of probability over a specified area of a joint distribution. If this information is used in the calculation, more accurate results can be obtained than that without using this information.

Another situation is uncertainty about the parameters of a distribution. All these topics are researched in this work. They are implemented in the software we call Statool.

Based on the developed methods, uncertainty can be flexibly considered and added into models. This can make the model closer to real situations. One problem posed by Sandia National Laboratory is studied in this work. Other applications include Pert networks, decision models and others.

# CHAPTER 1. INTRODUCTION

## Background

Uncertainty exists frequently in our knowledge of the real world. Consider two forms of uncertainty. One is variability coming from stochasticity, which is related to randomness. The other is epistemic uncertainty, also called other names as well. It comes from ignorance or imprecision. In principle, this kind of uncertainty can be reduced by additional empirical data (Ferson 2003).

Probability theory is a common approach to measuring the level of uncertainty. Probability density functions (PDFs) or their integrals, cumulative distribution functions (CDFs), are often used to model uncertainty in the value of a quantity. Often, uncertainty can be stated directly using a random variable. But this is not enough. People sometimes need to define random variables which are derived from arithmetic operations on other random variables. The distribution describing this random variable can be termed a derived distribution (Springer, 1979).

A variety of methods have been developed to compute derived distributions. Generally there are two classes of methods to handle it: analytical and numerical. Analytical methods are restricted to specific classes of input distribution, under assumptions such as independence. For example, normal distributions are often used. For examples, if two random variables are normal and independent, the sum of these two random variables is also normal. It is also possible to obtain derived distributions for specified dependency relationships other than independence, such as perfect positive rank correlation. However, it is often not easy or practical to find analytical results for random variable operations and it is not always a good idea to make convenient assumptions about dependency. Sometimes, we don't have any information about dependency. However, an advantage of analytical methods is accurate results. Unlike analytical methods, numerical methods only give numerical

results, but are suitable for a wide class of distributions. Numerical methods are widely used in real applications if approximate results can be accepted within specific tolerances.

Monte Carlo simulation is one of the best-known numerical methods. However, the traditional approach of Monte Carlo has some limitations. It assumes the distribution of the random variables is known, and their relationships are independent or known (Ferson 1996). If either the probability distributions or the dependency relationships of random variables are not available, some assumptions are usually made to fill in the missing knowledge. If the assumptions don't hold, results can be seriously affected. Thus we wish to be able to avoid such assumptions.

A discretized convolution approach can be used to calculate the result for the independent situation (Ingram et al. 1968; Colombo and Jaarsma 1980; Kaplan 1981). Interval analysis can be used to solve this problem. (It is obvious that intervals will be arbitrarily close to point values if the intervals are narrow enough.)

Intervals have the potential for bounding the result of an arithmetic operation on distributions. Discretization error coming from discretizing distributions may be bounded by interval based discretization (Berleant 1993). If the dependency is not specified, result bounds can be computed to include the entire range of possible dependencies. These bounds will generally be wider than if a particular dependency is specified. Interval-based dependency bounds analysis is described by Berleant and Goodman-Strauss (1998). This approach has fundamental similarities with the copula-based approach (Frank et al. 1987), which was significantly extended by Williamson and Downs (1990). These two methods have been implemented in software. The copula-based approach, termed probabilistic arithmetic, is implemented in the commercial software RiskCalc (Ferson et al. 1998). DEnv is implemented as Statool (Berleant and Goodman-Strauss 1998), which extends the previous tool (Berleant and Cheng 1998) by eliminating the independence assumption. Statool can handle the case where a dependency relationship is unknown or unspecified, by not making

any assumption about the dependency relationship between operands. But partial dependence information might be available in some cases. If we can use this information in the calculation, we will get more accurate results than can be obtained without using this information. Doing these is part of the research I have been involved in.

## The Distribution Envelopes Determination algorithm (DEnv): Interval-based analysis

An interval can be used to bound the range for a value. This interval may be associated with a specified probability, as when the domain of a random variable is partitioned. The partitioning of the domain of a random variable into intervals and probabilities is the basis for extending binary operations from intervals to distributions.

At this point, we only consider binary operations. We can however extend past binary operations and later we will talk about how to do this. Assuming there are 2 random variables $X$ and $Y$, to get the exact distribution for the result of an operation, we must know the joint distributions for random variables $X$ and $Y$. The joint distribution is related to the correlation of these two random variables. Here is an example.

Consider two random variables $X$ and $Y$. This table shows their distributions.

Table 1. Distributions for $X$ and $Y$.

| | $X$ | | | $Y$ | | |
|---|---|---|---|---|---|---|
| Range | [1,2] | [2,3] | [3,4] | [2,3] | [3,4] | [4,5] |
| Probability | 0.25 | 0.5 | 0.25 | 0.5 | 0.3 | 0.2 |

We don't have any information about distribution within particular intervals. And we also don't have any information about the dependency relationship between $X$ and $Y$. Therefore we don't know the joint distribution for $X$ and $Y$.

Consider addition: $Z=X+Y$. Because we don't have the joint distribution for $X$ and $Y$, it is impossible to find the exact result for $Z$. However we can put these two random variables into a matrix as shown in the following table, called a joint distribution tableau.

Table 2. Joint distribution tableau for $X$ and $Y$.

| | | | |
|---|---|---|---|
| $z \in [3,5]$<br>$p_{11} = ?$ | $z \in [4,6]$<br>$p_{12} = ?$ | $z \in [5,7]$<br>$p_{13} = ?$ | $y \in [2,3]$<br>$p_{Y1} = 0.5$ |
| $z \in [4,6]$<br>$p_{21} = ?$ | $z \in [5,7]$<br>$p_{22} = ?$ | $z \in [6,8]$<br>$p_{23} = ?$ | $y \in [3,4]$<br>$p_{Y2} = 0.3$ |
| $z \in [5,7]$<br>$p_{31} = ?$ | $z \in [6,8]$<br>$p_{32} = ?$ | $z \in [7,9]$<br>$p_{33} = ?$ | $y \in [4,5]$<br>$p_{Y3} = 0.2$ |
| $x \in [1,2]$<br>$p_{X1} = 0.25$ | $x \in [2,3]$<br>$p_{X2} = 0.5$ | $x \in [3,4]$<br>$p_{X3} = 0.25$ | $\leftrightarrow$ $\updownarrow$<br>$X$ $Y$ |

In Table 2, the last row in the table discretizes the distribution for $X$ and last column discretizes the distribution for $Y$. We don't know the values for probabilities $p_{11}$ through $p_{33}$ because we don't know the joint distribution, so question marks are shown. However, if $X$ and $Y$ are independent, we can fill in the missing values, as in the following table.

Table 3. Joint distribution for independency.

| | | | |
|---|---|---|---|
| $z \in [3,5]$<br>$p_{11} = 0.125$ | $z \in [4,6]$<br>$p_{12} = 0.25$ | $z \in [5,7]$<br>$p_{13} = 0.125$ | $y \in [2,3]$<br>$p_{Y1} = 0.5$ |
| $z \in [4,6]$<br>$p_{21} = 0.075$ | $z \in [5,7]$<br>$p_{22} = 0.15$ | $z \in [6,8]$<br>$p_{23} = 0.075$ | $y \in [3,4]$<br>$p_{Y2} = 0.3$ |
| $z \in [5,7]$<br>$p_{31} = 0.05$ | $z \in [6,8]$<br>$p_{32} = 0.1$ | $z \in [7,9]$<br>$p_{33} = 0.05$ | $y \in [4,5]$<br>$p_{Y3} = 0.2$ |
| $x \in [1,2]$<br>$p_{X1} = 0.25$ | $x \in [2,3]$<br>$p_{X2} = 0.5$ | $x \in [3,4]$<br>$p_{X3} = 0.25$ | $\leftrightarrow$ $\updownarrow$<br>$X$ $Y$ |

Thus, we can see that the joint distribution is affected by the dependency relationship between $X$ and $Y$. If we don't know the relationship between $X$ and $Y$, we can't determine the interior cell probabilities of a joint distribution tableau. But we can still infer some things about the result random variable $Z$ from this matrix. For example, consider sample value $z$=5. It only can occur in the grey cells of Table 4 as follow:

Table 4. Gray cells indicate possible cases of $z$=5.

| | | | |
|---|---|---|---|
| $z \in [3,5]$<br>$p_{11} = ?$ | $z \in [4,6]$<br>$p_{12} = ?$ | $z \in [5,7]$<br>$p_{13} = ?$ | $y \in [2,3]$<br>$p_{Y1} = 0.5$ |
| $z \in [4,6]$<br>$p_{21} = ?$ | $z \in [5,7]$<br>$p_{22} = ?$ | $z \in [6,8]$<br>$p_{23} = ?$ | $y \in [3,4]$<br>$p_{Y2} = 0.3$ |
| $z \in [5,7]$<br>$p_{31} = ?$ | $z \in [6,8]$<br>$p_{32} = ?$ | $z \in [7,9]$<br>$p_{33} = ?$ | $y \in [4,5]$<br>$p_{Y3} = 0.2$ |
| $x \in [1,2]$<br>$p_{X1} = 0.25$ | $x \in [2,3]$<br>$p_{X2} = 0.5$ | $x \in [3,4]$<br>$p_{X3} = 0.25$ | $\leftrightarrow$ $\updownarrow$<br>$X$ $Y$ |

We don't know the exact probability for $z \leq 5$. But we can consider the possible probabilities for $z \leq 5$. As Table 4 shows, only grey cells contribute to the probability of $z \leq 5$. We would like to determine the maximum value possible for this probability and the minimum also. To get the maximum value, all cells in which $Z$ can be $\leq 5$ will have their probabilities summed. To obtain the minimum value, only cells in which $Z$ must be $\leq 5$ will have their probabilities summed. For example, consider cell $p_{12}$. When we calculate the maximum value, the probability of this cell must be counted because $Z$ can be $\leq 5$ in the cell. But for the minimum value, we don't count this cell because $Z$ might not be $\leq 5$ in the cell. This way, we can find the possible range of cumulative probabilities for various values of $Z$. We can find the maximum probability and minimum probability for every value of $Z$ and connect all these points to get 2 curves: a maximum curve and a minimum curve. All the CDFs that are possible for $Z$ must be between these two curves.

In this example, possible value of $Z$ range is from 3 to 9. It is clear that the probability for $Z<3$ is zero and for $Z\leq 9$ is 1. The following part discusses the probability of $Z \leq 4$.

Maximum: We try to find all the cells in which $Z \leq 4$ *may* occur. From table 2, these cells are $p_{11}$, $p_{12}$, and $p_{21}$. So the maximum value will be the sum of $p_{11}$, $p_{12}$, and $p_{21}$.

Minimum: To obtain the minimum, we will find all the cells in which $Z$ *must* be $\leq 4$. In this table, there are none. Although $p_{11}$, $p_{12}$, and $p_{21}$ may satisfy $Z \leq 4$, they also might not. For example, the whole probability for each of those cells might be concentrated at the high bound of its range. So there is no cell in which $Z$ must be $\leq 4$.

Summarizing the above analysis, we can define a way to tell which cells contribute to the maximum and minimum probability values. All the cells in which the low bound is not greater than the value of $Z$ contribute to the maximum value. All the cells in which the high bound is not greater than the value of $Z$ contribute to the minimum value.

After finding all the cells satisfying the max (or min) condition, we will calculate the sum of the probabilities of these cells. Based on table 2, there exist constraints for the probabilities $p_{ij}$. It is clear that the sum of the $p_{ij}$'s in a row or column must equal the marginal probability of that row or column. These constraints can be described as follows:

Row Constraints: $\sum_{j=1}^{3} p_{ij} = p_{Yi}$ for i=1 to 3

Column Constraints: $\sum_{i=1}^{3} p_{ij} = p_{Xj}$ for j=1 to 3

Therefore, the questions become: find the maximum and minimum value for the sum of cells under these constraints. For the case $Z \leq 4$, these questions are as follows:

Maximum - make the sum of the specified cells' values as big as possible, that is, find

max ($p_{11}$+$p_{12}$+$p_{21}$)

such that:

$$\sum_{j=1}^{3} p_{ij} = p_{Yi} \text{ for } i=1 \text{ to } 3$$

and

$$\sum_{i=1}^{3} p_{ij} = p_{Xj} \text{ for } j=1 \text{ to } 3.$$

Minimum - make the sum of specified cells' values as small as possible, that is, find

$$\min \left( \sum_{i=1}^{3} \sum_{j=1}^{3} p_{ij} \right)$$

such that:

$$\sum_{j=1}^{3} p_{ij} = p_{Yi} \text{ for } i=1 \text{ to } 3$$

and

$$\sum_{i=1}^{3} p_{ij} = p_{Xj} \text{ for } j=1 \text{ to } 3.$$

For these two optimization questions, linear programming is a good tool to find the solution. LP allows us to find the probability range for the specified value of $Z$. The Table 5 shows the probabilities for various ranges of values for $Z$.

From the Table 5, we can draw two curves, a left curve and a right curve, using the maximum and minimum probabilities shown for $Z$. These two curves are called envelopes for the CDF of derived variable $Z$ because the CDF for derived variable $Z$ must be between these 2 curves whatever the dependency relationship between $X$ and $Y$ is. Figure 1 shows the final result.

Table 5. Cumulative probabilities for result variable $Z$, based on the joint distribution of Table 2 and linear programming.

| Z range | Maximum probability | Minimum probability |
|---|---|---|
| $p(Z<3)$ | 0 | 0 |
| $p(Z\leq3)$ | 0.25 | 0 |
| $p(Z\leq4)$ | 0.75 | 0 |
| $p(Z\leq5)$ | 1 | 0 |
| $p(Z\leq6)$ | 1 | 0.25 |
| $p(Z\leq7)$ | 1 | 0.55 |
| $p(Z\leq8)$ | 1 | 0.8 |
| $p(Z\leq9)$ | 1 | 1 |
| $p(Z\leq10)$ | 1 | 1 |

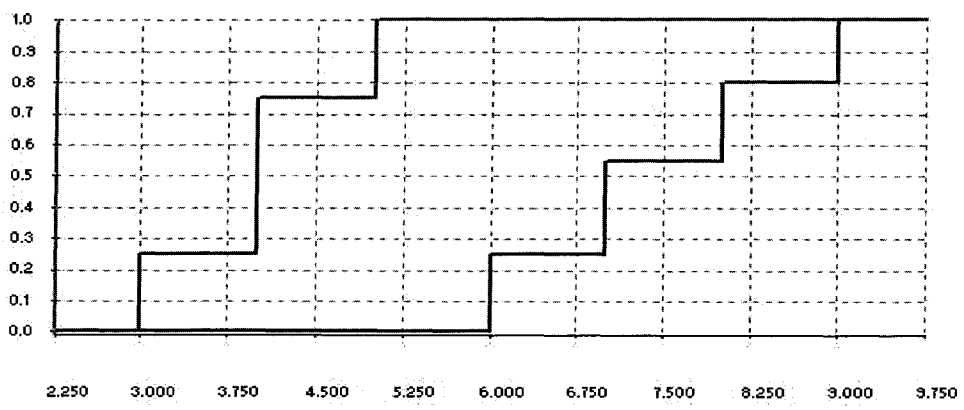



Figure 1. Probability bounds for $Z$.

## Objectives

The objectives of this dissertation research focus on the following situations.

### Using correlation as dependency information to improve results

Berleant and Goodman-Strauss (1998) addressed the situation in which a dependency relationship is unknown or unspecified by not making any assumption about the dependency relationship between operands. But partial dependence information might be available in some cases. If we can use this information in the calculation, we will get more accurate results than can be obtained without using this information. The objective here is to add this information into the computation to get an improved result. We consider Pearson correlation as partial dependency information later.

### CDFs for interval-parameterized distributions

We often have the situation that a distribution is undetermined due to not knowing the exact values of its parameter(s). Partial information about parameters may however be known. For example, the ranges of parameters may be given. A useful objective is to determine the envelope curves for such a distribution from the parameter range(s).

### Using partial information about the distribution of the result

Based on empirical data or design requirements, we may have partial information about the result, such as the known probability for a specified range of its support. This information may allow narrowing the envelope separation, and not only that associated with the specified range.

### Engineering applications

Using the research results to solve application problems.

### Implementing the methods in software

Statool is a useful tool for arithmetic on random variables. I added new functionalities to it to deal with partial information about dependency.

## Dissertation organization

The remainder of this dissertation is divided into 6 chapters and an appendix. The following 4 chapters are selected papers from scientific journals and proceedings. Each chapter focuses on one question and addresses one of the objectives. This means that each chapter can be read independently of the others.

The second chapter addresses how to use Pearson correlation information to improve the results. Usually the relationship between two operands is assumed to be independent. This assumption is not always true for applications. If the dependency relationship is considered instead as unknown, the independence assumption is removed, but the result is weakened. Instead there may some information about relationship. Pearson correlation is an example of partial information about the relationship between operands.

The chapter 3 considers interval parameters for common distributions. It provides methods to find the bounding envelopes for a CDF if the parameters are uncertain but in some range.

The fourth chapter presents other approaches to using partial information to improve results. The following situations are considered.

1. Knowledge about probabilities of specified areas of the joint distribution.
2. Knowledge about probabilities of specified ranges of values of the derived random variable.
3. Known relationships among the probabilities of different areas of the joint distribution, such as that the joint distribution is unimodal.
4. Known relationships among the probabilities of different ranges of the derived variable.

The fifth chapter uses the DEnv algorithm to solve a set of challenging problems posed in the literature: uncertainty in system response given uncertain parameters. Oberkampf et al. (2004) proposed a problem set which concentrates on the representation,

aggregation, and propagation of epistemic uncertainty and mixtures of epistemic and aleatory uncertainty through two simple model systems. Different investigators applied different methods to solve the problem (Ferson 2004).

General conclusions are presented in the last chapter of this dissertation.

Appendix contains new information about software, Statool. It is a useful tool for computing the algorithms described in this thesis. It is complete and easy to use, and is downloadable from the Web.

## Bibliography

[1] Alefeld, G. and J. Herzberger, Introduction to Interval Computations, Academic Press, New York, 1983.

[2] Berleant, D., Automatically verified reasoning with both intervals and probability density functions, Interval Computations (1993 No. 2), pp. 48-70.

[3] Berleant, D. and C. Goodman-Strauss, Bounding the results of arithmetic operations on random variables of unknown dependency using intervals, Reliable Computing 4(2) (1998), pp. 147-165.

[4] Berleant, D., L. Xie, and J. Zhang, Statool: a tool for distribution envelope determination (DEnv), an interval-based algorithm for arithmetic on random variables, Reliable Computing 9 (2) (2003), pp. 91-108.

[5] Berleant, D. and J. Zhang, Representation and problem solving with the Distribution Envelope Determination (DEnv) method, Reliability Engineering and System Safety 85 (2004) pp. 153-168.

[6] Berleant, D. and J. Zhang, Using correlation to improve envelopes around derived distribution, Reliable Computing 10 (2004), pp. 139-161.

[7] Berleant, D., The Thickets Approach to P-Bounds, Manuscript.

[8] Berleant, D., J. Zhang, R. Hu, and G. Sheblé, Economic dispatch: applying the interval-based distribution envelope algorithm to an electric power problem, SIAM Workshop on Validated Computing 2002 Extended Abstracts, Toronto, May 23-25, pp. 32-35.

[9] Berleant, D., L. Xie, J. Zhang, and G. Sheblé, An improved tool for distribution envelope determination, a technique for interval-based, verified arithmetic on random variables, SIAM Workshop on Validated Computing 2002 Extended Abstracts, Toronto, May 23-25, pp. 26-31.

[10] Berleant, D. and J. Zhang, Bounding the times to failure of 2-component systems, IEEE Transactions on Reliability, 53 (2004), pp. 542-550.

[11] Box, M.J., D. Davies, and W.H. Swann, Non-linear optimization techniques, 1969, Oliver & Body.

[12] Ferson, S., What Monte Carlo methods cannot do, Journal of Human and Ecological Risk Assessment 2 (4) (1996), pp. 990-1007.

[13] Ferson, S., V. Kreinovich, L. Ginzburg, D. Myers, and K. Sentz, Constructing Probability Boxes and Dempster-Shafer Structures, SAND REPORT SAND2002-4015, Sandia National Laboratories, Jan. 2003.

[14] Ferson, S., C. Joslyn, J. Helton, W. Oberkampf, and K. Sentz, Summary from the Epistemic Uncertainty Workshop: Consensus Amid Diversity. *Reliability Engineering and System Safety* 85 (2004), pp. 355-369.

[15] Hillier, F. S. and G. J. Lieberman, Introduction to operations research, McGraw-Hill, c2001

[16] Murty, K.G. and R. E. Krieger, Linear and combinatorial programming, Publishing Company, 1985.

[17] Neumaier, A., Clouds, fuzzy sets and probability intervals, Reliable computing 10 (2004), pp. 249-272.

[18] Oberkampf, W., J. Helton, C. Joslyn, S. Wojtkiewicz, and S. Ferson, Challenge problems: uncertainty in system response given uncertain parameters, Reliability Engineering and System Safety 85 (2004), pp. 11-19.

[19] Regan, H., S. Ferson, and D. Berleant, Equivalence of five methods for bounding uncertainty, Journal of Approximate Resoning, accepted pending revisions.

[20] Moore, R. E., Interval analysis, Prentice-Hall, Inc. 1966.

[21] Moore, R.E., Methods and applications of interval analysis, SIAM, 1979.

[22] Schach, S.R., Software engineering with Java, 1997, McGraw-Hill.

[23] Sheblé, G. and D. Berleant, Bounding the composite value at risk for energy service company operation with DEnv, an interval-based algorithm, SIAM Workshop on Validated Computing 2002, Extended Abstracts, Toronto, May 23-25, pp. 166-171.

[24] Siegrist, K., Virtual Laboratories in Probability and Statistics. Web site http://www.math.uah.edu/statold.

[25] Smith, J.E., Generalized Chebychev Inequalities: theory and application in decision analysis, Operations Research, (1995) 43: 807-825.

[26] Springer, M.D., The algebra of Random variables, Wiley, 1979.

[27] Tucker, W.T. and S. Ferson, Probability bounds analysis in environmental risk assessments, Applied Biomathematics, 2003.

[28] Walsh, G.R., Methods of optimization, 1975, John Wiley & Sons.

[29] Williamson, R. and T. Downs, Probabilities arithmetic I: numerical methods for calculating convolutions and dependency bounds, International Journal of Approximate Reasoning 4 (1990).

[30] Zhang, J. and D. Berleant, Envelopes around cumulative distribution functions from interval parameters of standard continuous distributions, Proceedings, North American Fuzzy Information Processing Society (NAFIPS 2003), Chicago, pp. 407-412.

# CHAPTER 2. USING PEARSON CORRELATION TO IMPROVE ENVELOPES AROUND THE DISTRIBUTIONS OF FUNCTIONS

A paper published in the Journal of Reliable Computing 10: 139-161, 2004.

Daniel Berleant and Jianzhong Zhang

## Abstract

Given two random variables whose dependency relationship is unknown, if a new random variable is defined whose samples are some function of samples of the given random variables, the distribution of this function is not fully determined. However, envelopes can be computed that bound the space through which its cumulative distribution function must pass. If those envelopes could be made to bound a smaller space, the cumulative distribution, while still not fully determined, would at least be more constrained. We show how information about the correlation between values of given random variables can lead to better envelopes around the cumulative distribution of a function of their values.

## Introduction and background

A random variable whose samples are a function of samples of other random variables is often called a derived random variable and its distribution *a derived distribution.* Given two random variables with samples $u$ and $v$, probability density functions $f_u(.)$ and $f_v(.)$, and cumulative distribution functions $F_u(.)$ and $F_v(.)$, a sample $x$ of a derived distribution can be defined in various ways, such as:

- $x=u+v$ (Frank et al. [11]);
- $x=\max(u,v)$, which models the time to complete two concurrent tasks; and

- $x_v = (38u - 8v)/(0.08u + 0.048v)$, where u and v are the fuel cost rates of two electric generators and $x_v$ is the optimal power output of the generator with rate $v$ (Wood and Wollenberg [22]).

We wish to describe the distribution $F_x(.)$ of $x$.

Derived distributions may be determined analytically or numerically. Analytical methods tend either to assume distributions are of particular forms or, in the case of moment propagation, to ignore other information about the distributions. Springer [18] gives a reasonably comprehensive account up to its time of publication. We pursue the numerical strategy here. Our strategy represents each input probability density function (PDF) discretely using a histogram-like set of intervals with associated probabilities [2]. The discretized inputs form the marginals of a discretized joint distribution termed a *joint distribution tableau.* Each cell in a joint distribution tableau contains an interval and a probability, and is termed a marginal cell if it contains an interval $\boldsymbol{u}_i$ or $\boldsymbol{v}_j$ in the discretization of marginal $f_u(.)$ or $f_v(.)$, and an interior cell if contains an interval and a probability $p_{ij}$ (Table 1). For each $i, j$, $p_{ij} = p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j)$. If the inputs are statistically independent in the usual sense then $p_{ij} = p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j) = p(u \in \mathbf{u}_i) \times p(v \in \mathbf{v}_j)$.

Table 1. a joint distribution tableau. Independent random variables with PDFs $f_u(.)$ and $f_v(.)$ are shown in discretized form, using intervals $u_1$, $u_2$, and $u_3$ and their associated probabilities to represent $f_u(.)$ in the bottom row, and intervals $v_1$, $v_2$, and $v_3$ and their associated probabilities to represent $f_v(.)$ in the left column. Values $u$ and $v$ are drawn from $f_u(.)$ and $f_v(.)$. The discretization is coarse for illustration. We have discretized $f_u(.)$ and $f_v(.)$ without overlaps, so some intervals have open endpoint(s). These are shown with a parenthesis instead of a square bracket.

| | | | |
|---|---|---|---|
| $\mathbf{v}_3$=(5,9] | $x$=(2.5,9] | $x$=(5/3,9/2] | $x$=(5/4,3] |
| $p(v \in \mathbf{v}_3) = 0.1$ | $p_{13}$=0.02 | $p_{23}$=0.05 | $p_{33}$=0.03 |
| $\mathbf{v}_2$=(4,5] | $x$=(2,5] | $x$=(4/3,5/2] | $x$=(1,5/3] |
| $p(v \in \mathbf{v}_2) = 0.8$ | $p_{12}$=0.16 | $p_{22}$=0.4 | $p_{32}$=0.24 |
| $\mathbf{v}_1$=[0,4] | $x$=[0,4] | $x$=[0,2) | $x$=[0,4/3) |
| $p(v \in \mathbf{v}_1) = 0.1$ | $p_{11}$=0.02 | $p_{21}$=0.05 | $p_{31}$=0.03 |
| $v\uparrow$ $x=v/u$ | $\mathbf{u}_1$=[1,2] | $\mathbf{u}_2$=(2,3] | $\mathbf{u}_3$=(3,4] |
| $u\rightarrow$ | $p(v \in \mathbf{u}_1) = 0.2$ | $p(v \in \mathbf{u}_2) = 0.5$ | $p(v \in \mathbf{u}_3) = 0.3$ |

If each probability $p_{ij}$ in Table 1 is assumed to be distributed uniformly over its corresponding interval $\boldsymbol{v}_j/\boldsymbol{u}_i$, a not unreasonable approximation if the discretization is sufficiently fine, then the cumulative distribution function (CDF) of $x$, call it $F_x(x_0)$, could be plotted by taking values $x_0$ and performing the following steps for each (Moore [13]).

1. Integrate each interior cell from $-\infty$ to $x_0$.
2. Sum the integrals computed for the interior cells.

The PDF $f_x(.)$ instead of the CDF $F_x(.)$ can also be obtained (Ingram et al. 1968; Colombo and Jaarsma 1980). If no assumption is made about the distribution of the $p_{ij}$'s over their respective domains, then $F_x(.)$ cannot be determined precisely, but can be bounded with envelopes (Figure 1) which bound the effects of discretization [2].

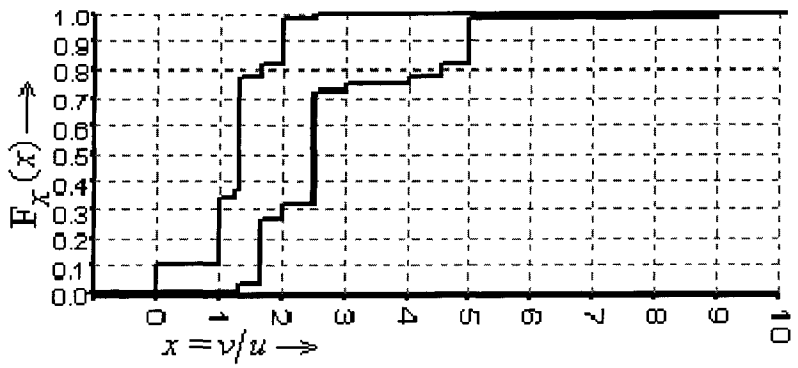



Figure 1. envelopes $F_x(x)$ around CDF $F_x(x)$, where $x=v/u$ and $f_u(.)$ and $f_v(.)$ are discretized as shown in Table 1.

A problem with such methods is the need to know the dependency relationship between the input distributions. Independence is a common assumption in practice though not always justified. Independence as well as other dependency relationships (as in Table 2) can be represented in a joint distribution tableau by appropriate choice of interior cell probabilities. However, sometimes no specification of dependency is justified by what is known about the problem.

Table 2. a joint distribution tableau like that of Table 1 except with different values for the $p_{ij}$'s, indicating that the joint distribution is different. Hence the dependency relationship between values $u$ and $v$ of the marginals is also different.

| | | | |
|---|---|---|---|
| $\mathbf{v}_3$=(5,9] | $x$=(2.5,9] | $x$=(5/3,9/2] | $x$=(5/4,3] |
| $p(v \in \mathbf{v}_3) = 0.1$ | $p_{13}$=0.1 | $p_{23}$=0 | $p_{33}$=0 |
| $\mathbf{v}_2$=(4,5] | $x$=(2,5] | $x$=(4/3,5/2] | $x$=(1,5/3] |
| $p(v \in \mathbf{v}_2) = 0.8$ | $p_{12}$=0 | $p_{22}$=0.5 | $p_{32}$=0.3 |
| $\mathbf{v}_1$=[0,4] | $x$=[0,4] | $x$=[0,2) | $x$=[0,4/3) |
| $p(v \in \mathbf{v}_1) = 0.1$ | $p_{11}$=0.1 | $p_{21}$=0 | $p_{31}$=0 |
| $v\uparrow$ $x=v/u$ | $\mathbf{u}_1$=[1,2] | $\mathbf{u}_2$=(2,3] | $\mathbf{u}_3$=(3,4] |
| $u\rightarrow$ | $p(v \in \mathbf{u}_1) = 0.2$ | $p(v \in \mathbf{u}_2) = 0.5$ | $p(v \in \mathbf{u}_3) = 0.3$ |

There are a number of approaches to the problem of numerically computing derived distributions without specifying a dependency relationship between the operands (Figure 2 shows an example). One is Monte Carlo simulation (MC), as in Red-Horse and Benjamin [5]. However the randomness inherent in MC can lead to complications in the results and their interpretation (Ferson 1996). Another approach is based on copulas (Frank et al. 1987; Nelsen 1999), and a tool implementing the Probabilistic Arithmetic (Williamson and Downs 1990) extension of Frank et al. is available commercially (Ferson 2002). A third approach, clouds, was recently proposed by Neumaier [5]. A fourth approach is discrete convolution of the actual distributions. Various techniques based on this approach have existed since at least as early as 1968 [5] for the case of independence. More recently, the technique described here, called Distribution Envelope Determination (DEnv), extended the discrete convolution technique to the case of unknown dependency (Berleant and Goodman-Strauss 1998 [5]). DEnv is reviewed in the next section. It is implemented in a downloadable tool called Statool [5].

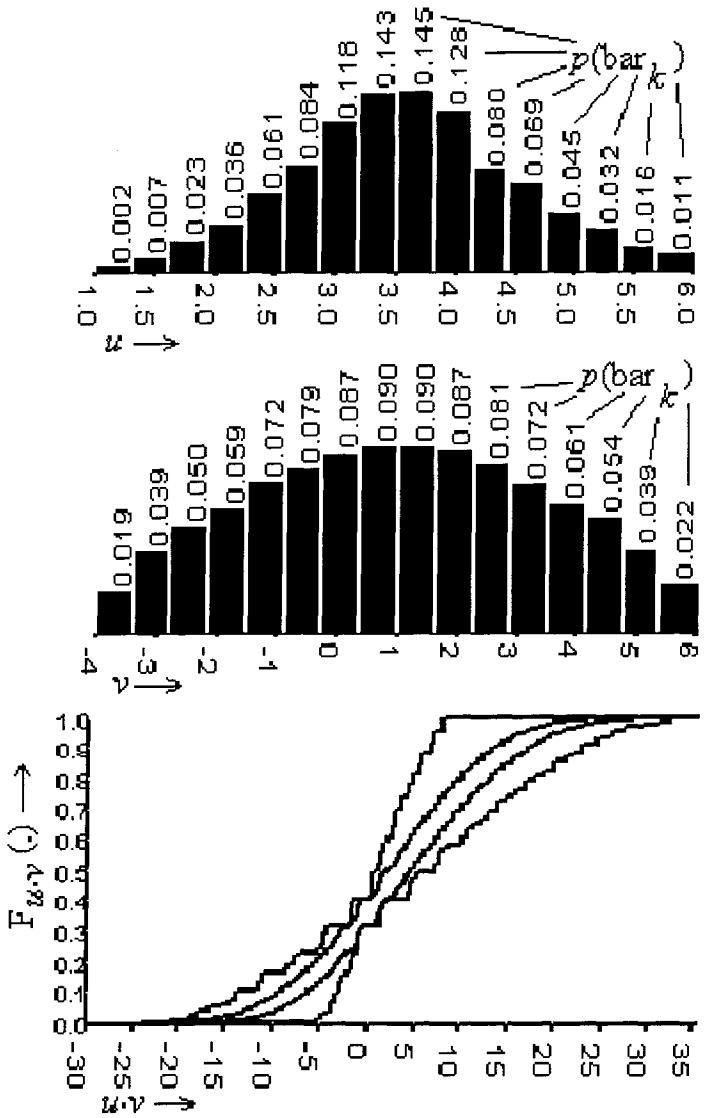



Figure 2. (*top and middle*) histogram-like discretizations of input PDFs $f_u(u)$ and $f_v(v)$. Each bar is labeled at the top with its probability. (*Bottom*) two pairs of envelopes around $F_{u.v}(.)$, the CDF of derived value $x=u.v$. The two exterior envelopes bound the CDF when the dependency relationship between $u$ and $v$ is unknown. The two interior envelopes bound the CDF when $u$ and $v$ are independent. In the independent case, the envelopes are non-identical because they bound the effects of information loss due to discretization. The rougher appearance of both pairs of envelopes near $u.v=0$ is because 0·anything=0.

Finally, the intermediate situation of *partial information* about the dependency may occur. There is a need for ways to use partial information about dependency between inputs when determining envelopes around the CDFs of derived distributions [5]. A common and important way to express partial information about dependency is correlation. Correlation constitutes partial information because it does not fully characterize a dependency relationship (different joint distributions can have exactly the same correlation). We have extended DEnv to incorporate information about correlation. We use Pearson correlation, the most common kind and the kind normally implied by uses of the otherwise ambiguous term "correlation." (In copula-based approaches, handling Pearson correlation is problematic [5] because converting joint distributions into copulas involves stretching the marginals into a normalized form, and Pearson correlation depends on the un-normalized forms.) The purpose of this paper is to report on an extension of DEnv that uses Pearson correlation as a problem input.

We review the DEnv algorithm next (a more detailed account appears in [5]). Then we explain how to extend DEnv to use correlation to provide constraints that can often decrease the separation of the envelopes.

## Distribution Envelope Determination (DEnv): a review

**The goal.** DEnv obtains boundaries around the space through which a derived CDF may travel (Figure 2). More specifically, let $F_x(.)$ be the cumulative distribution for $x$, where $x$ is a function of $u$ and $v$. The density function $f_u(.)$ of $u$ is discretized with a set of intervals $\mathbf{u}_i$, each associated with a probability such that the sum of these probabilities is 1. Density function $f_v(.)$ of $v$ is similarly discretized with a set of intervals $\mathbf{v}_j$. Because the discretizations lose information that is present in the undiscretized $f_u(.)$ and $f_v(.)$, there will typically not be a single CDF that is implied for $x=g(u,v)$ even when the dependency relationship is fully specified [5]. Our objective then is to obtain left and right envelopes around the family of

CDFs that are possible for $x$. These envelopes may be expressed symbolically as the interval-valued function $\mathbf{F}_x(.)$. The left (top) envelope then is $\overline{\mathbf{F}_x}(.)$ and the right (bottom) envelope is $\underline{\mathbf{F}x}(.)$

**The givens.** Envelope computation takes as input the correlation between the marginals, when that is available, and a joint distribution tableau. A joint distribution tableau discretely represents a family of joint distributions containing all joint distributions that are consistent with that discretization. For example, recall the joint distribution tableau of Table 1. This tableau states that $p(v \in [0,4]) = 0.1$, $p(u \in [1,2]) = 0.2$, and $p(v \in [0,4] \cap u \in [1,2]) = 0.02$. Each cell in the tableau contains an interval-valued bin in which $u$ or $v$ (for a marginal cell), or $x=v/u$ (for an interior cell) might fall, and a probability that it falls in that bin. The probabilities of interior cells are specified if the dependency relationship of the marginals is known, and not specified if the dependency is not known. There are many variations in how values $u$ and $v$ of the marginals can be distributed, and in how they can be jointly distributed, that are consistent with these bin specifications. Put another way, Table 1 gives a correct discretization of any pair of marginal distributions and their joint distribution for which the statements in all of the cells are correct. Table 1 also contains a discretization of the distribution of $x$. This is the set of interior cells, each of which specifies an interval-valued bin for $x=v/u$ and a probability $p_{ij}$.

In the following subsections we first assume independence in the traditional sense (Section 1), then extend that to arbitrary dependency relationships (Section 2), then further the algorithm to the case of an unknown dependency relationship (Section 3). With that as background the case of a dependency relationship constrained by correlation is finally addressed (Section 3). That case constitutes the new contribution of this report.

## Solution for independent marginals

Equations (1)-(2) summarize the solution for the general case of $x=g(u,v)$, with interval extension $\mathbf{x}_{ij}=\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)$ where $\mathbf{u}_i$ and $\mathbf{v}_j$ are intervals in discretizations of the distributions $f_u(.)$ and $f_v(.)$ from which values $u$ and $v$ are drawn.

$$\overline{\mathbf{F}_x}(x_0) = \sum_{i,j:\underline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)}\le x_0} p(u \in \mathbf{u}_i)\cdot p(v \in \mathbf{v}_j), \qquad (1)$$

$$\underline{\mathbf{F}_x}(x_0) = \sum_{i,j:\overline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)}\le x_0} p(u \in \mathbf{u}_i)\cdot p(v \in \mathbf{v}_j). \qquad (2)$$

The summations are over all pairs $i$, $j$ such that $\underline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0$ in Equation (1), or $\overline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0$ in Equation (2).

We first explain why Equation (1) computes the left bounding envelope $\overline{\mathbf{F}_x}(x)$, using an example. Then the differences for the right envelope are noted. Bolding will indicate an interval and overlining the upper bound of an interval.

### Computing the left (upper) envelope from a joint distribution tableau

The example is based on Table 1 and is stated in several steps.

For x<0, $\overline{\mathbf{F}_x}(x) = 0$ because no interior cell contains an interval containing any values below zero, so p(x<0) must be zero.

For 0≤x≤1, $\overline{\mathbf{F}_x}(x) = p_{11} + p_{21} + p_{31} = 0.1$ for the following reasons.

1. Only the interior cells containing $p_{11}$, $p_{21}$, and $p_{31}$ have intervals with low bounds ≤1. Therefore only those cells can contain $x$ when x≤1, thereby contributing their probability to the cumulative probability $\overline{\mathbf{F}_x}(x)$.
2. Call the distribution of a particular interior cell's probability over its interval its mini-distribution. The probability associated with an interior cell must be distributed somehow within its interval, but *mini-distributions* are not

otherwise defined. Therefore to obtain the height of the left bounding envelope at a given value of $x$ we must assume that each mini-distribution has a form that leads to the greatest possible height at that value. The simplest such assumption is that the mini-distribution of each interior cell interval is an impulse at its low bound, because then each interior cell whose interval low bound is at or below a value $x$=$x_0$will contribute all of its probability to $\overline{\mathbf{F}_x}(x_0)$.

For 1<x≤5/4, p32=0.24 can also contribute to $\overline{\mathbf{F}_x}(x)$, so $\overline{\mathbf{F}_x}(x) = p_{11} + p_{21} + p_{31} + p_{32} = 0.34$

For 5/4<x≤4/3, p33=0.03 can also contribute to $\overline{\mathbf{F}_x}(x)$, for a total cumulative probability of 0.37.

This line of reasoning continues until all interior cells contribute their probabilities to $\overline{\mathbf{F}_x}(x)$, resulting in the staircase-shaped left envelope shown in Figure 1.

### Computing the right (lower) envelope

The right bounding envelope, $\underline{\mathbf{F}_x}(x)$, is derived similarly, except that points on it are obtained by assuming that the probability in each interior cell is an impulse at its interval high bound instead of its interval low bound.

## Solution for an arbitrary dependency between the marginals

In Table 1, each $p_{ij} = p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j)$ is the product of the probabilities of its corresponding marginal cells. This is consistent with the traditional definition of statistical independence. Other assignments of probabilities to the $p_{ij}$'s imply other dependency relationships. If the dependency relationship is known then the interior cells can be filled in so that their probabilities are consistent with that dependency relationship and its joint distribution. In such cases the value of each $p_{ij}$ is not necessarily $p_{ij} = p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j)$, instead arising out of the dependency relationship, which defines the value

of $p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j)$. This implies a generalization of Equations (1)-(2), shown as Equations (3)-(4).

$$\overline{\mathbf{F}_x}(x_0) = \sum_{i,j:\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)\le x_0} p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j), \qquad (3)$$

$$\underline{\mathbf{F}_x}(x_0) = \sum_{i,j:\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)\le x_0} p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j). \qquad (4)$$

## Solution for the case of unknown dependency between the marginals

As explained earlier (Section 2), the interior cells of a joint distribution tableau represent a family of CDFs. When the dependency relationship between the marginals is unknown, then Equations (1)-(4) cannot be evaluated because the $p_{ij}$'s are not determined. Intuitively, because the $p_{ij}$'s are now variable, they may take on values consistent with a greater variety of joint distributions, and hence a greater variety of CDFs for derived random variable $x$. This will tend to make the envelopes bounding this larger family of CDFs wider apart. An augmentation to the algorithm is required to deal with this situation. The augmented algorithm is described next in two steps, one short and one longer, and then summarized in Equations (5)-(9).

1. *Determine which interior cells contribute.* The same cells contribute their probabilities to the CDF at a value of $x$ as would contribute in the case of known dependency, and for the same reasons. These are the cells specified by Equations (1)-(4).
2. *Maximize (for the left envelope), or minimize (for the right envelope) the sum of the probabilities of the contributing cells.* Because the $p_{ij}$'s are not fully determined when the dependency relationship is unknown, DEnv finds maximums and minimums given the result of step 3 by manipulating the $p_{ij}$'s in the joint distribution tableau. Call the interior cells identified in step 3 the

contributing cells, and the cells containing the remaining $p_{ij}$'s the non-contributing cells.

To maximize the sum of the probabilities of the contributing cells, we transfer as much probability as possible from non-contributing interior cells to contributing interior cells. To illustrate, recall Table 1. If the assumption that the marginals are independent is relaxed, the $p_{ij}$'s are underdetermined. However they are constrained by the fact that the probabilities of the interior cells in any given row must sum to the probability of its corresponding marginal cell, and similarly for any given column (Table 3).

Table 3. (*top*) a joint distribution tableau like that of Tables 1 and 2, but showing only the $p_{ij}$'s and without values assigned to them. (*Bottom*) the constraints that the tableau defines on the values of the $p_{ij}$'s. Each constraint states that the sum of the probabilities of the $p_{ij}$'s in a row or column equals the probability in the marginal cell for that row or column. This follows from standard properties of joint distributions and their marginals.

| | | | |
|---|---|---|---|
| $p(v \in \mathbf{v}_3) = 0.1$ | $p_{13}$ | $p_{23}$ | $p_{33}$ |
| $p(v \in \mathbf{v}_2) = 0.1$ | $p_{12}$ | $p_{22}$ | $p_{32}$ |
| $p(v \in \mathbf{v}_1) = 0.1$ | $p_{11}$ | $p_{21}$ | $p_{31}$ |
| $v\uparrow \quad u\rightarrow$ | $p(u \in \mathbf{u}_1) = 0.2$ | $p(u \in \mathbf{u}_2) = 0.5$ | $p(u \in \mathbf{u}_3) = 0.3$ |

| **Row constraints** | **Column constraints** |
|---|---|
| $p_{11}+p_{21}+p_{31}=0.1$ | $p_{11}+p_{12}+p_{13}=0.2$ |
| $p_{12}+p_{22}+p_{32}=0.8$ | $p_{21}+p_{22}+p_{23}=0.5$ |
| $p_{13}+p_{23}+p_{33}=0.1$ | $p_{31}+p_{32}+p_{33}=0.3$ |

For example, compare the assignment of probability to $p_{32}$ in Table 2 with its assignment in Table 1, 0.3 vs. 0.24. In Table 2, $\overline{\mathbf{F}}_x(1.1) = p_{11} + p_{21} + p_{31} + p_{32} = 0.4$, which is greater than the 0.34 implied by $p_{11}+p_{21}+p_{31}+p_{32}$ in Table 1. $\overline{\mathbf{F}}_x(1.1)$ can be no higher than

the Table 2 value of 0.4 no matter what the joint distribution is, because the third row must comply with the constraint $p_{11} + p_{21} + p_{31} = p(v \in \mathbf{v}_1) = 0.1$, and the only contributing interior cell outside of the third row is the one containing $p_{32}$, which can be no higher than 0.3 because its column must comply with the constraint $p_{31} + p_{32} + p_{33} = p(u \in \mathbf{u}_3) = 0.3$. The result is a point on the left envelope $x$=1.1 that is higher than the envelope derived for the independent case, a new height that applies not only to $x=v/u$=1.1 but also to all values of $x=v/u$ for which the contributing cells are $p_{11}, p_{21}, p_{31}$, and $p_{32}$. For other values of $x$ the set of contributing cells is different, so the $p_{ij}$'s of Table 2 might not lead to the highest possible value of $\overline{\mathbf{F}}_x(x)$. In that case some other set of assignments of probabilities to the $p_{ij}$'s consistent with Table 3 will result in the highest possible value instead. Thus for each value of $x=v/u$ it is necessary to find the contributing cells, and assignments to the $p_{ij}$'s in them that lead to the highest possible value of $\overline{\mathbf{F}}_x(x)$. The result is ultimately a left envelope that is farther to the left than the left envelope shown in Figure 1. Similar reasoning based on minimization instead of maximization gives a new right envelope that is farther to the right than the one shown in Figure 1.

Maximizing the collective probability of a set of contributing cells by the ad hoc reasoning process used for $x$=1.1 for various values of $x$ would rapidly become tedious to do manually. Fortunately a general and automatable method is available in the form of linear programming (LP). LP optimizes (maximizes or minimizes) a linear function, called the objective function, with respect to a set of linear constraints. The linear function to optimize in this case is the sum of the probabilities of the contributing cells. LP will maximize this consistently with the linear constraints imposed by the marginals, one constraint for each $\mathbf{u}_i$ and one for each $\mathbf{v}_j$ in the joint distribution tableau (Table 3). LP is invoked and its output, the maximum (minimum) possible total probability that can be allocated among the contributing cells, is the $y$ coordinate associated with $x$, thus completing the coordinates for a point on the left (right) envelope.

The extensions of Equations (1)-(2) and (3)-(4) to objective functions to optimize for the unknown dependency situation are:

$$\overline{\mathbf{F}_x}(x_0) = \max \sum_{i,j:\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)\le x_0} p_{ij} \qquad (5)$$

for the left envelope, and

$$\underline{\mathbf{F}_x}(x_0) = \min \sum_{i,j:\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)\le x_0} p_{ij} \qquad (6)$$

for the right envelope. The applicable constraints are:

$$\sum_j p_{ij} = p(\mathbf{u}_i), \qquad \text{for all } i, \qquad (7)$$

$$\sum_i p_{ij} = p(\mathbf{v}_j), \qquad \text{for all } j, \qquad (8)$$

$$p_{ij} \ge 0, \qquad \text{for all } i,j. \qquad (9)$$

## Using correlation to move the envelopes closer together

Specifying a dependency relationship between the input random variables implies envelopes that are closer together than when the dependency is unknown (Figure 2). A value or range for correlation is a *partial* specification of the dependency, and so implies envelopes that are:

- at least as close together as when the dependency is unknown, but
- at least as far apart as when the dependency is fully specified.

DEnv infers the effects of constraints on envelopes via calls to a linear programming routine. Thus to use information about correlation, this information must be expressed as linear constraints. These constraints can then supplement the row and column constraints used by the LP calls. This is explained next, while Section 4 provides examples.

We begin with a standard formula for the Pearson correlation ρ. We use Pearson correlation in this paper as it is the most common kind of correlation and is usually implied by otherwise unqualified uses of the term "correlation."

$$\rho = \frac{\mathbf{E}(uv) - \mathbf{E}(u)\mathbf{E}(v)}{\sqrt{\left[\mathbf{E}(u^2) - \mathbf{E}(u)^2\right]\left[\mathbf{E}(v^2) - \mathbf{E}(v)^2\right]}} = \frac{\mu_{u.v} - \mu_u \cdot \mu_v}{\sqrt{\sigma_u^2 \cdot \sigma_v^2}} \qquad (10)$$

Here ρ is the Pearson correlation coefficient of the distributions of *u* and *v*, *u* and *v* are values to be drawn from the marginal distributions, **E**(*u*) is the expectation function and is equivalent to the mean $\mu_u$, $\mathbf{E}(u^2) - \mathbf{E}(u)^2 = \sigma_u^2$ is the variance of *u*, and similarly for *v*. Since ρ and the marginals are problem inputs, all terms can be computed from the inputs except **E**(*uv*), the only term that depends on the joint distribution. Solving for **E**(*uv*) gives

$$\mathbf{E}(uv) = \mathbf{E}(u)\mathbf{E}(v) + \rho\sqrt{\left[\mathbf{E}(u^2) - \mathbf{E}(u)^2\right]\left[\mathbf{E}(v^2) - \mathbf{E}(v)^2\right]} \qquad (11)$$

Because DEnv uses the PDFs of *u* and *v* after they have been discretized into sets of intervals and their associated probabilities, and because the distribution of each associated probability over its interval is unspecified, terms in Equation (11) can be determined only to within intervals. For example, given the discretized distribution of *v* in Tables 1 and 2,

$$\mathbf{E}(v) \in 0.1 * [0,4] + 0.8 * (4,5] + 0.1 * (5,9] = (3.7,5.3] \qquad (12)$$

If we follow the convention of bolding interval-valued symbols, then **E**(*v*)=(3.7,5.3]. This leads to an intervalized form of Equation (11) suitable for use with discrete representations of PDFs for *u* and *v*, and interval constraints on ρ.

$$\mathbf{E}_g = \mathbf{E}(uv) = \mathbf{E}(u)\mathbf{E}(v) + \rho\sqrt{\left[\mathbf{E}(u^2) - \mathbf{E}(u)^2\right]\left[\mathbf{E}(v^2) - \mathbf{E}(v)^2\right]} = \mu_u\mu_v + \rho\sigma_u^2\sigma_v^2 \qquad (13)$$

Thus **E**(*uv*) is calculated from **ρ** and discretizations of the PDFs of *u* and *v*. Since **ρ** and the marginals are **g**ivens, we will call this expectation $\mathbf{E}_g$.

Another way to calculate **E**(*uv*) is directly from a joint distribution tableau. This gives an interval for **E**(*uv*), namely $\sum_{ij} \mathbf{u}_i \mathbf{v}_j p_{ij}$ . See Table 4. Because it is computed as a property of the joint distribution, as expressed discretely by a given joint distribution **t**ableau, call it $\mathbf{E}_t(.)$. Its argument is a joint distribution tableau with a fully specified set of value assignments to the $p_{ij}$'s. The assignment of probability values to the interior cells of the joint distribution tableau, in conjunction with the $\mathbf{u}_i\mathbf{v}_j$ intervals, implies an interval $\mathbf{E}_t(.)$ that must

be consistent with $\mathbf{E}_g$ (which represents the discretized distributions of $u$ and $v$ and the given correlation). If $\mathbf{E}_t(.)$ and $\mathbf{E}_g$ are not consistent with each other, that assignment of values to the $p_{ij}$'s is not consistent with the given correlation and therefore is not allowed. As the following steps show, consistency means that $\mathbf{E}_t(.)$ and $\mathbf{E}_g$ overlap.

1. $\mathbf{E}_g$ is the interval of admissible values for $\mathbf{E}(uv)$ based on ρ and other problem inputs as specified in Equation (13). The terms in (13) are all calculated from the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s (see e.g. Equation (12)). Because the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s appear repeatedly in (13), naïve interval evaluation will often result in $\mathbf{E}_g$ containing excess width, thereby weakening the power of $\mathbf{E}_g$ as a constraint on admissible values of $\mathbf{E}(uv)$. To avoid that, an optimization technique can be used to compute good bounds for $\mathbf{E}_g$. Alternatively, values or ranges for the means ($\mu_u$ and $\mu_v$) and variances ($\sigma_u^2$ and $\sigma_v^2$) of the marginals can be provided as problem inputs. This has the added benefit of allowing incorporation of mean and variance information that may be available and more specific than the bounds for mean and variance derivable directly from the discretized marginals.

2. $\mathbf{E}_t(.)$, in contrast to $\mathbf{E}_g$, is affected by the $p_{ij}$'s, which are determined by the joint distribution. An expression for $\mathbf{E}_t(.)$ may be derived as follows.

$$\overline{\mathbf{E}_t(.)} = \overline{\sum_{i,j} \mathbf{u}_i \mathbf{v}_j p_{ij}} = \sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j p_{ij}} = \sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j} p_{ij} \quad (14)$$

$$\underline{\mathbf{E}_t(.)} = \sum_{i,j} \underline{\mathbf{u}_i \mathbf{v}_j} p_{ij}$$

To compute bounds on $\mathbf{E}_t(.)$ using Equations (14), the numerical value of each $\underline{\mathbf{u}_i \mathbf{v}_j}$ and $\overline{\mathbf{u}_i \mathbf{v}_j}$ term is needed. The standard definition of interval multiplication accounts for all possible combinations of signs on the bounds of $\mathbf{u}_i$ and $\mathbf{v}_j$ by multiplying each bound of $\mathbf{u}_i$ by

each bound of $\mathbf{v}_j$ (four combinations), and using the *min* and *max* of the four as $\underline{\mathbf{u}_i\mathbf{v}_j}$ and $\overline{\mathbf{u}_i\mathbf{v}_j}$ respectively (e.g. Alefeld and Herzberger 1983).

Table 4. abstract template for joint distribution tableaus. The bottom row includes a marginal cell describing the case where a value $u$ drawn from marginal $f_u(.)$ falls within interval $u_i$ of the discretization of $f_u(.)$. The left column includes a similar cell for $v$, $f_v(.)$ , and $v_j$. The function for combining values $u$ and $v$ is $g(u,v)=x$, its interval extension is $g(u_i,v_j)=x_{ij}$, and the distribution of value $x=g(u,v)$ is represented discretely by the interior cells of the tableau, one of which is shown in detail. Product $u_iv_j$ is used in calculating $E_t(.)$, which is the range of possible values of $E(uv)$ for the tableau.

| ..... | ..... | ..... | ..... |
|---|---|---|---|
| $\mathbf{v}_j$ | ..... | $x = g(u,v) \in \mathbf{g}(\mathbf{u}_i,\mathbf{v}_j) = \mathbf{x}_{ij}$<br>$p_{ij} = p(u \in \mathbf{u}_i \cap v \in \mathbf{v}_j)$<br>$\mathbf{u}_i\mathbf{v}_j = [\min(\underline{\mathbf{u}_i}\underline{\mathbf{v}_j}, \underline{\mathbf{u}_i}\overline{\mathbf{v}_j}, \overline{\mathbf{u}_i}\underline{\mathbf{v}_j}, \overline{\mathbf{u}_i\mathbf{v}_j}),$<br>$\max(\underline{\mathbf{u}_i}\underline{\mathbf{v}_j}, \underline{\mathbf{u}_i}\overline{\mathbf{v}_j}, \overline{\mathbf{u}_i}\underline{\mathbf{v}_j}, \overline{\mathbf{u}_i\mathbf{v}_j})]$ | ..... |
| ..... | ..... | ..... | ..... |
| $v\uparrow$ $x=g(u,v)$<br>$u\rightarrow$ | ..... | $\mathbf{u}_i$ | .... |

3. The $p_{ij}$'s are variables because they are under-determined by the row and column constraints (Table 3). Assigning a specific set of values to the $p_{ij}$'s implies an associated interval $\mathbf{E}_t(.)$, which can be calculated per Equations (14). Some sets of value assignments to the $p_{ij}$'s imply intervals for $\mathbf{E}_t(.)$ that do not overlap $\mathbf{E}g$. Those assignments are inconsistent with the correlation provided as a problem input (as explained in detail in the next step), and so can be excluded as implausible.

Excluding a set of assignments to the $p_{ij}$'s can move the left envelope toward the right of where it would be if there was no information about correlation, and/or move the right envelope toward the left, narrowing their separation. This is because the excluded set of assignments might have a higher maximum cumulation $\overline{\mathbf{F}_x(x)}$ or lower minimum cumulation $\underline{\mathbf{F}_x(x)}$ for a given value of $x$ than any that are not excluded.

4. The previous step stated that $\mathbf{E}_t(.)$ and $\mathbf{E}g$ are inconsistent when they have no overlap. This step explains why. Specifying the values of the $p_{ij}$'s does not define the distribution of any $p_{ij}$ over $\mathbf{u}_i\mathbf{v}_j$. Hence a joint distribution tableau with specified values for its $p_{ij}$'s represents a *family* of joint distributions. All joint distributions that conform to the discretization expressed by the joint distribution tableau are in that family.

A joint distribution for values $u$ and $v$ has a numerical value for $\mathbf{E}(uv)$. $\mathbf{E}_t(.) = \sum_{i,j} \mathbf{u}_i \mathbf{v}_j p_{ij}$ thus gives the range of numerical values for $\mathbf{E}(uv)$ exhibited by the various joint distributions in the family associated with a particular set of value assignments to the $p_{ij}$'s. If $\mathbf{E}_t(.)$ does not intersect $\mathbf{E}g$, then there is no joint distribution in that family for which $\mathbf{E}(uv) \in \mathbf{E}_g$, so that set of value assignments to the $p_{ij}$'s is excludable as inconsistent with the value $\rho$ or range $\boldsymbol{\rho}$ provided as a problem input. This requirement that $\mathbf{E}_t(.)$ and $\mathbf{E}g$ overlap is stated in inequality form as the following two constraints:

$$\underline{\mathbf{E}_g} \le \overline{\mathbf{E}_t(.)} \text{ and } \overline{\mathbf{E}_g} \ge \underline{\mathbf{E}_t(.)} \qquad (15)$$

5. To use constraints (15) in a linear programming problem, symbols $\underline{\mathbf{E}_g}$ and $\overline{\mathbf{E}_g}$ are replaced with their numerical values as calculated in step 13. $\underline{\mathbf{E}_t(.)}$ and $\overline{\mathbf{E}_t(.)}$ are replaced with $\sum_{i,j} \underline{\mathbf{u}_i \mathbf{v}_j p_{ij}}$ and $\sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j p_{ij}}$ respectively, as described in step 13. This results in Equations (16).

$$\underline{\mu_u \mu_v + \rho\sqrt{\sigma_u^2 \sigma_v^2}} \le \sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j} p_{ij} \text{ and } \overline{\mu_u \mu_v + \rho\sqrt{\sigma_u^2 \sigma_v^2}} \ge \sum_{i,j} \underline{\mathbf{u}_i \mathbf{v}_j} p_{ij} \qquad (16)$$

Since the only variables in Equations (16) are the $p_{ij}$'s, (16) constitutes linear constraints as required by LP. These can supplement the row and column constraints (Table 3), and will tend to result in envelopes that are closer together than those resulting from the row and column constraints alone.

## Strengthening the effect of correlation

The width of interval $\mathbf{E}_t(.) = \sum_{i,j} \mathbf{u}_i \mathbf{v}_j p_{ij}$ is derived from the widths of the $\mathbf{u}_i \mathbf{v}_j$ terms. However if the distribution of each probability $p_{ij}$ over the corresponding interval $\mathbf{u}_i \mathbf{v}_j$ was fully defined then the overall distribution of $uv$ would be fully defined. Then a numerically-valued function, call it $E_t(.)$, could be calculated instead of the interval-valued function $\mathbf{E}_t(.)$. To define the distribution of each $p_{ij}$ one might consider assuming that, as examples, the distribution of each $p_{ij}$ over the interval $\mathbf{u}_i \mathbf{v}_j$ is uniform, or is an impulse at the midpoint of $\mathbf{u}_i \mathbf{v}_j$, or has some other fully defined form.

Since $E_t(.)$ is a number it will be narrower than the interval $\mathbf{E}_t(.)$, unless $\mathbf{E}_t(.)$ is a thin interval containing only one number. (This will occur in the important special case where $u$ and $v$ are discretized as series of impulses.) Suppose $E_t(.)$ is in fact narrower. Then it is less likely to intersect with $\mathbf{E}_t(.)$ and so more likely to be excluded as inconsistent with $\mathbf{E}_g$. Thus constraints (15) would be strengthened, leading to envelopes that are closer together.

For example, assume the distribution of each $p_{ij}$ is uniform over $\mathbf{u}_i \mathbf{v}_j$. Since the expectation of a uniform distribution is its midpoint, Equations (14) become

$$\overline{\mathbf{E}_t(.)} = \underline{\mathbf{E}_t(.)} = E_t(.) = \sum_{i,j} mid(\mathbf{u}_i \mathbf{v}_j) \cdot p_{ij} \qquad (17)$$

where *mid(.)* is the midpoint of its interval argument. Then (15) becomes the stronger pair of constraints

$$\underline{\mathbf{E}_g} \le E_t(.) \text{ and } \overline{\mathbf{E}_g} \ge E_t(.) \qquad (18)$$

The effect of correlation can be strengthened not only by narrowing $\mathbf{E}_t(.)$, but also by narrowing $\mathbf{E}_g$. A way to narrow $\mathbf{E}_g$ is to accept as inputs point value(s) for expectations and variances $\mu_u=E(u)$, $\mu_v=E(v)$, $\sigma_u^2 = \left[E(u^2) - E(u)^2\right]$, and/or $\sigma_v^2 = \left[E(v^2) - E(v)^2\right]$, instead of calculating intervals for them from the discretized marginals as in step 13 of Section 3. If these were all point values then the width of $\mathbf{E}_g$ would be controlled by the width of $\boldsymbol{\rho}$, and if $\rho$ was a number then $\mathbf{E}_g$ would be a number (call it $E_g$) as well.

Since narrowing either $\mathbf{E}_t(.)$ or $\mathbf{E}_g$ tends to strengthen the effects of correlation, a third approach that narrows both is to use a finer discretization for the marginals. Finer discretizations narrow $\mathbf{E}_g$ by narrowing $\mathbf{E}(u)$, $\mathbf{E}(v)$, $\mathbf{E}(u^2)$ and $\mathbf{E}(v^2)$ in Equation (13), and also narrow computations of $\mathbf{E}_t(.)$ by narrowing the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s, resulting in narrower $\mathbf{u}_i\,\mathbf{v}_j$ terms in Equations (14). Other ways of expressing partial information about dependency, including identification of useful assumptions besides correlation, and when those assumptions are reasonable to make, are likely to enable additional progress in narrowing envelopes around derived distributions.

## Examples

We start with an example that is simple enough to go through in full detail, followed by another example of more realistic complexity.

### A basic, detailed example

Let the distribution for value $u$ consist of two impulses of equal probability: $\mathbf{u}_1$=[1,1] and $\mathbf{u}_2$=[100,100], with $p(u \in \mathbf{u}_1) = p(u \in \mathbf{u}_2) = 0.5$, and let the distribution describing $v$ be the same as for $u$. The joint distribution tableau is shown in Table 5. First the envelopes for the case of unknown dependency are derived. Then correlation is added as a constraint and we show how this reduces the separation between the envelopes.

Table 5. (*top*) joint distribution tableau for a simple problem. (*Bottom*) the linear constraints implied by the tableau.

| | | |
|---|---|---|
| $\mathbf{v}_2$=[100,100] | $u+v$=[101,101] | $u+v$=[200,200] |
| $p$=0.5 | $p_{12}$=? | $p_{22}$=? |
| $\mathbf{v}_1$=[1,1] | $u+v$=[2,2] | $u+v$=[101,101] |
| $p$=0.5 | $p_{11}$=? | $p_{21}$=? |
| $u+v$ | $\mathbf{u}_1$=[1,1] | $\mathbf{u}_2$=[100,100] |
| | $p$=0.5 | $p$=0.5 |

| **Constraint name** | **Equation** |
|---|---|
| Top row | $p_{12}+p_{22}=0.5$ |
| 2nd row | $p_{11}+p_{21}=0.5$ |
| 2nd column | $p_{12}+p_{11}=0.5$ |
| Right Column | $p_{22}+p_{21}=0.5$ |

## Unknown dependency condition

The left envelope may be derived as follows.

- For $u+v<2$, $\overline{\mathbf{F}_{u+v}(.)}=0$ because $u+v$ cannot be below 2.
- For $u+v\in[2,101)$, only $p_{11}$ contributes its probability to $\overline{\mathbf{F}_{u+v}(.)}$, and its maximum possible value is 0.5. This is because $p_{11}$=0.5 is consistent with the row and column constraints, shown in Table 5, by setting $p_{11}= p_{22}$=0.5 and $p_{12}= p_{21}$=0, while any value for $p_{11}$ over 0.5 would immediately violate the 2nd row and 2nd column constraints. Thus $\overline{\mathbf{F}_{u+v}(.)}=0.5$ in this case.
- For $u+v\in[101,200)$, $p_{11}$, $p_{12}$, and $p_{21}$ contribute to $\overline{\mathbf{F}_{u+v}(.)}$. Their sum $p_{11}+p_{12}+p_{21}$ can be as high as 1 while remaining consistent with the row and column constraints, by setting $p_{12}=p_{21}$ =0.5 and $p_{11}= p_{22}$=0. Thus $\overline{\mathbf{F}_{u+v}(.)}=1$ in this case.
- For $u+v\geq200$, $\overline{\mathbf{F}_{u+v}(.)}=1$ because $u+v$ must be at or below 200.

The right envelope may be derived as follows.

- For $u+v<2$, $\underline{\mathbf{F}_{u+v}(.)}=0$ because $u+v$ cannot be below 2.
- For $u+v\in[2,101)$, only $p_{11}$ contributes to $\underline{\mathbf{F}_{u+v}(.)}$. The minimum possible

value of $p_{11}$ is 0 because the row and column constraints are all satisfied if we set $p_{11}$= $p_{22}$=0 and $p_{12}$=$p_{21}$ =0.5. Thus $\underline{\mathbf{F}_{u+v}(.)} = 0$ in this case.

- For $u+v \in [101,200)$, $p_{11}$, $p_{12}$, and $p_{21}$ contribute to $\underline{\mathbf{F}_{u+v}(.)}$. Their sum $p_{11}$+ $p_{12}$+$p_{21}$ can be as low as 0.5 while remaining consistent with the row and column constraints, by setting $p_{12}$=$p_{21}$ =0 and $p_{11}$= $p_{22}$=0.5. Any value below 0.5 for the sum would immediately violate the 2nd row and 2nd column constraints. Thus $\underline{\mathbf{F}_{u+v}(.)} = 0.5$ in this case.
- For $u+v \geq 200$, $\underline{\mathbf{F}_{u+v}(.)} = 1$ because $u+v$ must be at or below 200.

The envelopes are shown in Figure 3. Next we show how correlation narrows the separation of these envelopes.

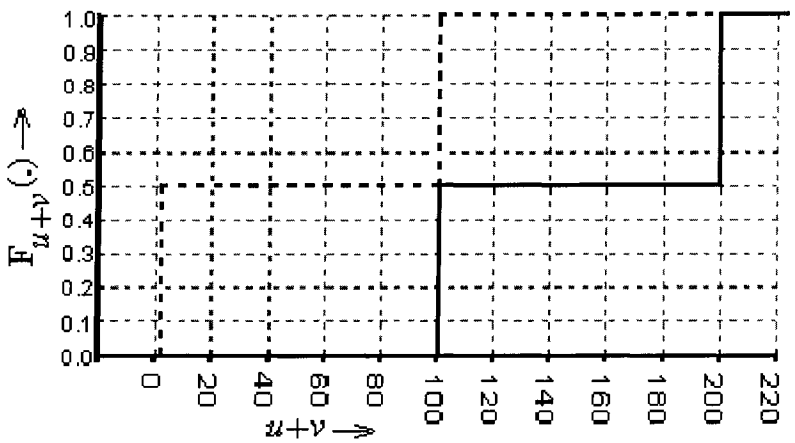



Figure 3. envelopes around the CDF of $u+v$, for the joint distribution tableau of Table 5.

## Effect of correlation

Let us illustrate how correlation works step by step, extending the example just detailed by incorporating the information that $\rho \in [0.7,1]$. From this, and the joint distribution tableau of Table 5, $\mathbf{E}_g$ may be calculated by substituting intervals into Equation (13) as follows:

$$\mathbf{E}_g = \frac{[1,1]+[100,100]}{2} \cdot \frac{[1,1]+[100,100]}{2} + [0.7,1] \cdot \sqrt{\left(\frac{[1,1]^2+[100,100]^2}{2} - \left(\frac{[1,1]+[100,100]}{2}\right)^2\right) \cdot \left(\frac{[1,1]^2+[100,100]^2}{2} - \left(\frac{[1,1]+[100,100]}{2}\right)^2\right)}$$

$$= [4265.425, 5000.5]$$

Next, values are substituted from the interior cells of the joint distribution tableau of Table 5 into Equation (14) to get an expression for $\mathbf{E}_t(.)$, as follows.

$$\mathbf{E}_t(.) = p_{11} \cdot [1,1] \cdot [1,1] + p_{12} \cdot [1,1] \cdot [100,100] + p_{21} \cdot [100,100] \cdot [1,1] + p_{22} \cdot [100,100] \cdot [100,100]$$
$$= p_{11} + 100p_{12} + 100p_{21} + 10000p_{22}$$

Thus $\mathbf{E}_t(.)$ is a thin interval in this example. To signify that, we will consider it a number and use the symbol $E_t(.)$ henceforth. The four constraints of Table 5 are augmented with the following two new constraints derived from the computations for $\mathbf{E}_g$ and $\mathbf{E}_t(.)$ just shown, and from Equations (15).

$$4265.425 \le p_{11} + 100p_{12} + 100p_{21} + 10000p_{22} \qquad (19)$$

$$5000.5 \ge p_{11} + 100p_{12} + 100p_{21} + 10000p_{22} \qquad (20)$$

**Applying the new constraints.** One can now ask how adding Constraints (19)-(20) to the row and column constraints leads to envelopes that are closer together than for the unknown dependency condition.

The new left envelope may be derived as follows.

- For $u+v<2$, the earlier conclusion, $\overline{\mathbf{F}_{u+v}(.)} = 0$, is unaffected.
- For $u+v \in [2,101)$ the earlier conclusion, $\overline{\mathbf{F}_{u+v}(.)} = 0.5$, occurs for $p_{11}= p_{22}=0.5$ and $p_{12}= p_{21}=0$, is unchanged because those assignments to the $p_{ij}$'s imply $\mathbf{E}_t(.) = 0.5 + 100 \cdot 0 + 100 \cdot 0 + 10000 \cdot 0.5 = 5000.5$, and 5000.5 is consistent with Constraints (19)-(20).
- For $u+v \in [101,200)$ the analysis is more involved. The earlier conclusion based on only the row and column constraints was that $\overline{\mathbf{F}_{u+v}(.)} = p_{11} + p_{12} + p_{21} = 1$ and that this could be achieved by setting $p_{12}=p_{21}=0.5$ and $p_{11}=p_{22}=0$. For the present scenario of $\rho \in [0.7,1]$, however, this result is too high because those assignments to the $p_{ij}$'s lead to the following calculation.

$$\mathbf{E}_t(.) = 1 \cdot 0 + 100 \cdot 0.5 + 100 \cdot 0.5 + 10000 \cdot 0 = 100 \qquad (21)$$

which violates Constraint (19). The reason is that these assignments to the $p_{ij}$'s allocate all the probability for value $u+v$ in Figure 5 to $p_{12}$ and $p_{21}$, which are in the cells for which one marginal has value 1 and the other has value 100. Thus when a value of one marginal is low the value of the other is high. This allocation is inconsistent with the given correlation of [0.7,1] which, being positive, requires $u$ and $v$ to tend to be either both low or both high.

To calculate a new value of $\overline{\mathbf{F}_{u+v}(.)}$ for $u+v \in [101,200)$ given $\rho \in [0.7,1]$, we can derive and solve simultaneous equations on the $p_{ij}$'s by hand or, as Statool does, invoke linear programming on a computer. For illustration we do it next using simultaneous equations.

The extreme of assigning all probability to $p_{12}$ and $p_{21}$ and no probability to $p_{11}$ and $p_{22}$, which gave the envelope height calculated earlier for the unknown dependency condition, is not possible for $\rho \in [0.7,1]$ as shown by Equation (21). We wish to reduce the sum $p_{11}+p_{12}+p_{21}$ (hence increasing $p_{22}$) just enough to raise $\mathbf{E}_t(.)$ from 100 up to 4265.425, because this will result in the maximum possible assignment to $p_{11}+p_{12}+p_{21}$ that is consistent with $\mathbf{E}_t(.) = p_{11} + 100p_{12} + 100p_{21} + 10000p_{22} \in [4265.425,5000.5]$, as required by Constraints (19)-(20). To do this we use, as one of the simultaneous equations, $p_{11} + 100p_{12} + 100p_{21} + 10000p_{22} = 4265.425$. Solving this simultaneously with the constraint equations of Table 5 gives $p_{11}+p_{12}+p_{21}=0.425+0.075+0.075=0.575$..

The conclusion is that, for $u+v \in [101,200)$ and $\rho \in [0.7,1]$, the left envelope height $\overline{\mathbf{F}_{u+v}(.)}$ is 0.575, which is considerably lower than its value of 1 under the unknown dependency condition.

6. For $u+v \geq 200$, the earlier conclusion that $\overline{\mathbf{F}_{u+v}(.)} = 1$ is unaffected.

The new right envelope may be derived as follows.

- For $u+v<2$, the earlier conclusion, $\underline{\mathbf{F}_{u+v}(.)} = 0$, is unaffected.

- For $u+v \in [2,101)$, only $p_{11}$ contributes to $\underline{\mathbf{F}_{u+v}(.)}$. The minimum possible value of 0 for $p_{11}$ found for the unknown dependency condition is too low for $\rho \in [0.7,1]$. This is because $p_{11}$=0 implies $p_{22}$=$p_{11}$=0 and $p_{12}$=$p_{21}$=0.5 due to the constraints shown in Table 5, just as in the discussion of $u+v \in [101,200)$ for the left envelope, above. There, we moved as small as possible an amount of probability out of $p_{11}$+ $p_{12}$+$p_{21}$, which was the sum of the contributing cell probabilities. This is the same as moving as small a probability as possible into $p_{22}$, the only non-contributing cell and thus the complement of the contributing cells. Here, we wish to move as small as possible an amount into $p_{11}$, not $p_{22}$, but because the constraints in Table 5 imply $p_{11}$= $p_{22}$, the resulting allocation of probabilities among the interior cells is actually the same. Thus, as above, Constraints (19)-(20) in conjunction with the constraints of Table 5 imply a minimum value for $p_{11}$= $p_{22}$ of 1-($p_{11}$+$p_{12}$+$p_{21}$)=*1-0.575=0.425*. Therefore for $u+v \in [2,101)$, when $\rho \in [0.7,1]$, $\underline{\mathbf{F}_{u+v}(.)} = 0.425$. This is considerably higher than its value of 0 under the unknown dependency condition.
- For $u+v \in [101,200)$, the earlier conclusion that $\underline{\mathbf{F}_{u+v}(.)} = 0.5$ occurs for $p_{12}$= $p_{21}$ =0 and $p_{11}$= $p_{22}$ =0.5 is unchanged, because those assignments to the $p_{ij}$'s imply $E_t(.) = 0.5 + 100 \cdot 0 + 100 \cdot 0 + 10000 \cdot 0.5 = 5000.5$, which is consistent with Constraints (19)-(20).
- For $u$+$v$≥200 the earlier conclusion, $\underline{\mathbf{F}_{u+v}(.)} = 1$, is unaffected.

The envelopes around the CDF of $u$+$v$ when $\rho \in [0.7,1]$ are shown in Figure 4. They are closer together than for the unknown dependency condition shown in Figure 3. For ease

of exposition the example just described used a joint distribution tableau containing numbers (or strictly speaking, thin intervals). If the marginal intervals are widened, giving weaker specifications for the inputs, wider envelopes around the CDF of $u+v$ result (Figure 5).

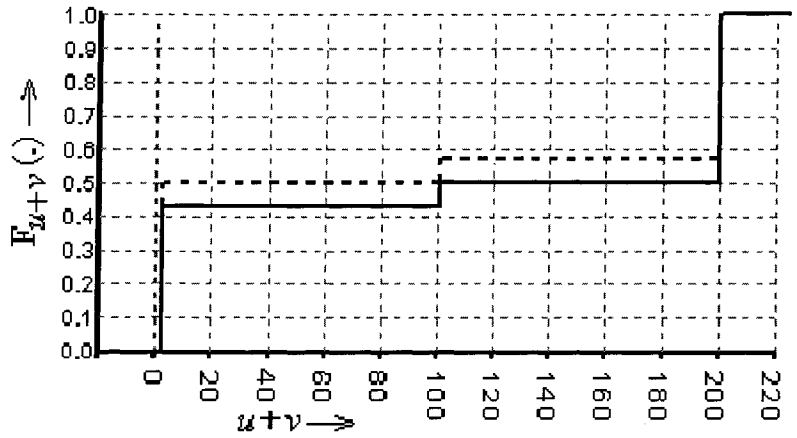



Figure 4. envelopes around the CDF of $u+v$, for the joint distribution tableau of Table 5, given $\rho \in [0.7,1]$.

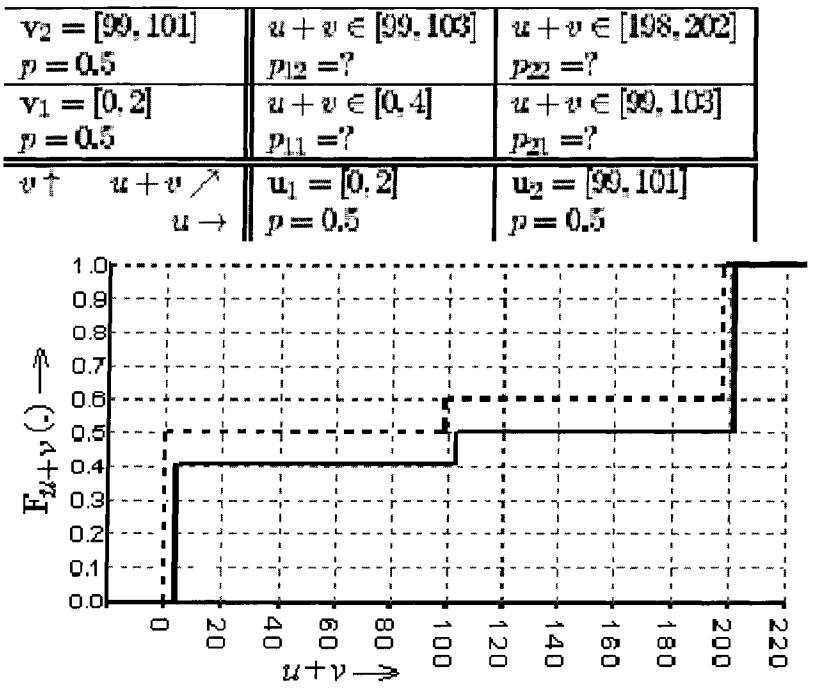



Figure 5. (*top*) joint distribution tableau like that of Table 5 except that the intervals are wider. (*Bottom*) envelopes around the CDF of $u+v$ for the joint distribution tableau at top, assuming $\rho \in [0.7,1]$. These envelopes are wider than the envelopes in Figure 4 because the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s here specify the PDFs for $u$ and $v$ more weakly, with widths of 2 instead of 0 as in Table 5.

## A more complex example

Here we show the effects of different correlation conditions using inputs with realistically detailed discretizations. Figures 6 and 7 show two discretized distributions. Let $u$ and $v$ be values drawn from the skewed distribution and the bimodal distribution, respectively. (Bimodal distributions can find application in describing system parameters that are controlled to stay within an allowable range. As the parameter wanders within this range, it often approaches the endpoints of the range, activating a control mechanism that prevents it from passing those endpoints. As a result the parameter may tend to spend more time near

the endpoints of the range than in the middle.) We further specify that $\mu_u \in [3,3.1]$, $\mu_v \in [5.15,5.25]$, $\sigma_u^2 \in [5,5.1]$, and $\sigma_v^2 \in [11.4,11.5]$.

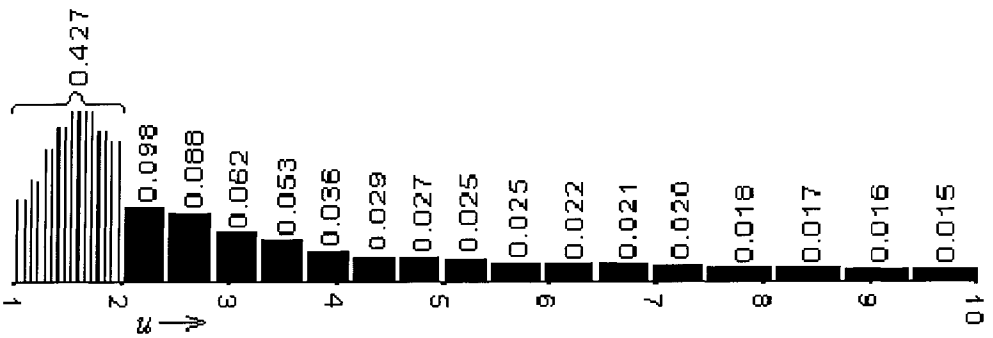



Figure 6. a discretized input distribution. The flat tops of the bars are an artifact of the graphical representation and do not imply uniform (or any other) distribution of probability over the domain of any given bar.

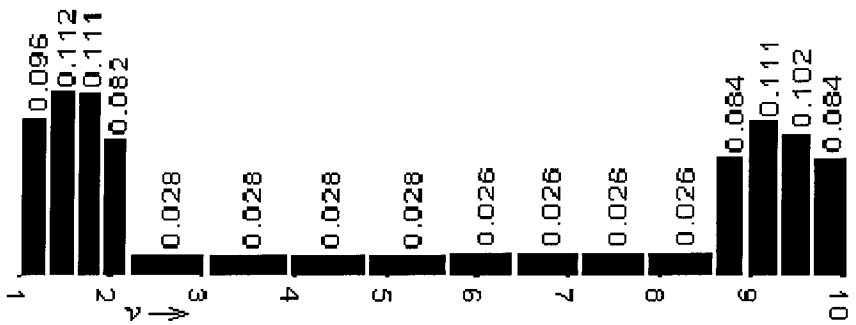



Figure 7. discretization of a bimodal PDF to be used as a divisor.

Let $z = \frac{u}{v}$, Assuming that values $u$ and $v$ are independent gives envelopes for $z$ that are relatively close together (Figure 8). The relatively small separation between them occurs because the algorithm automatically bounds the effects of discretization as noted in Section 2. Removing the independence assumption leads to envelopes that are much wider apart (Figure 9).

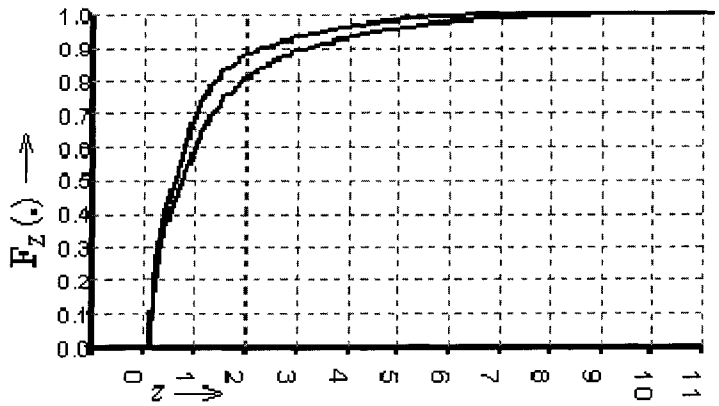



Figure 8. envelopes around the cumulative distribution for $z$, where $z = u/v$ and $u$ and $v$ are assumed independent. This is a strong assumption that leads to envelopes that are relatively close together.

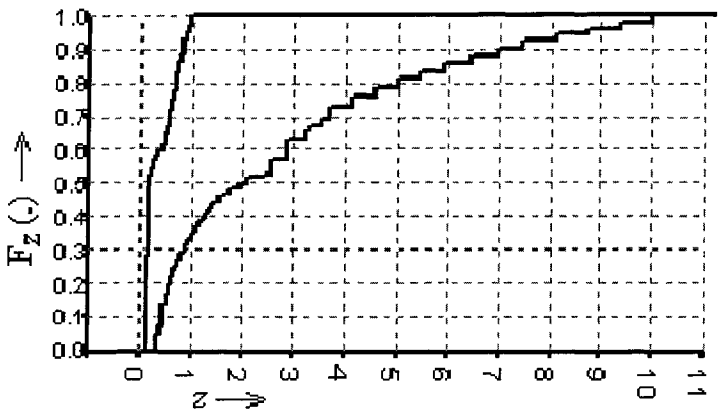



Figure 9. envelopes around the CDF of $z$, where $z = u/v$ and no assumptions are made about the dependency relationship between $u$ and $v$. The lack of information about dependency yields envelopes around the cumulative distribution of $z$ that are relatively widely separated.

Specifying that the correlation is negative, that is, that $\rho \in [-1,0)$, results in envelopes that are slightly narrower (Figure 10) than for the unknown correlation condition. Note for example the rounding of the northwest knee of the left envelope relative to the unknown correlation case in Figure 9. This rounding means we can, for example, rule out the

possibility that the CDF has value 1 (i.e. certainty) for some values on the horizontal axis, which could potentially be significant for decision-making. Restricting the sign of the correlation appears to usually be a rather weak constraint, since many different dependency relationships can have correlation measures with the same sign.

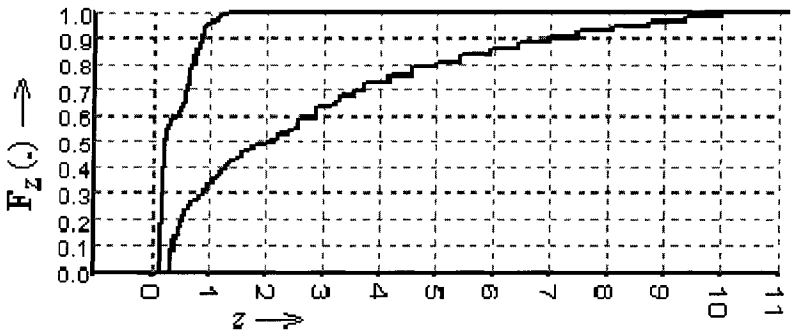



Figure 10. envelopes around the CDF of $z = u/v$, where $u$ and $v$ are assumed to have negative correlation ($\rho<0$).

Stronger correlations can have greater effects. Figure 11 shows 3 pairs of envelopes superposed. Progressing from weaker to stronger restrictions on correlation, the outermost envelopes bound the possible CDFs for $z$ given $\rho \in [-1,-0.5)$. The 2nd envelope from the left and 2nd envelope from the left bound the possible CDFs given [-1,-0.8]. The innermost envelopes bound the possible CDFs given the strongest restriction on correlation, $\rho \in [-1,-0.83)$.

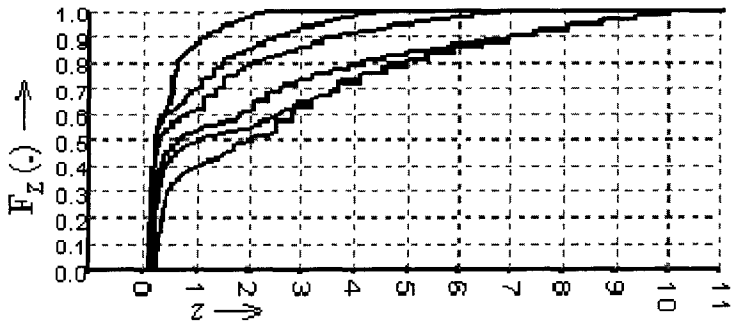



Figure 11. envelopes around the CDF for $z = u / v$ under three correlation conditions. The outermost envelopes are for the weakest of the three, $\rho \in [-1,-0.5]$. The envelopes 2nd from the left and 2nd from the right are for $\rho \in [-1,-0.8])$. The innermost envelopes are for the strongest correlation condition, $\rho \in [-1,-0.83)$.

## Conclusion

DEnv (Distribution Envelope Determination) is a numerical algorithm for computing envelopes around the space of possible cumulative distribution functions of derived random variables. These are random variables whose values are a function of the values of other random variable(s). Envelopes are appropriate for safely bounding the CDFs of derived random variables when the dependency relationship between the input distributions is not fully known. This is important because often available information is insufficient to reliably justify a particular dependency relationship. Each possible dependency relationship implies some CDF in the family that is bounded by the envelopes. Envelopes can also bound the effects of discretization, which occurs because DEnv requires that input distributions be discretized.

We have previously reported how DEnv can handle the case where the dependency relationship between input distributions is unknown. However, partial information about dependency may be available in the form of values or ranges for correlation. This paper extends the DEnv algorithm to incorporate such information about correlation. Pearson

correlation is used because it is the most commonly used kind of correlation. Some examples are provided, showing how correlation can strengthen results relative to those obtained without any information about dependency.

## Acknowledgements

The authors thank Gerald Sheblé for discussions regarding needs for advances in the DEnv algorithm and possibilities for its application. Using applications to drive advances in the theory and software for DEnv is an important part of our research strategy. As a result we continue to pursue applications in such areas as competitive bidding [5], financial engineering [5], and electric power generation [5].

The anonymous referees contributed significantly to the exposition. Referee #1 offered comprehensive suggestions for revision which are greatly appreciated. The authors retain full responsibility for any remaining shortcomings. This work has been supported in part by research funding from the Power Systems Engineering Research Center (PSERC).

## References

1. Alefeld, G. and Herzberger, J.: *Introduction to Interval Computations,* Academic Press, New York, 1983.
2. Berleant, D.: Automatically Verified Reasoning with Both Intervals and Probability Density Functions, *Interval Computations* (1993), pp. 48-70.
3. Berleant, D. and Goodman-Strauss, C.: Bounding the Results of Arithmetic Operations on Random Variables of Unknown Dependency Using Intervals, *Reliable* Computing 4 (2) (1998), pp. 147-165.
4. Berleant, D. and Zhang, J.: Representation and Problem Solving with the Distribution Envelope Determination (DEnv) Method, Reliability Engineering and System Safety, forthcoming. http://www.public.iastate.edu/~berleant/.

5. Berleant, D., Zhang, J., Hu, R., and Sheblé, G.: Economic Dispatch: Applying the Interval-Based Distribution Envelope Algorithm to an Electric Power Problem, *SIAM Workshop on Validated Computing 2002 Extended Abstracts*, Toronto, May 23-25, pp. 32-35. http://www.public.iastate.edu/~berleant/.
6. Cheong, M.-P.: *Competitive Bidding to Sell Power Under Epistemic Uncertainty About the Competition*, Master's thesis, Dept. of Electrical and Computer Engineering, Iowa State University, in preparation.
7. Colombo, A.G. and Jaarsma, R.J.: A Powerful Numerical Method to Combine Random Variables, *IEEE Transactions on Reliability* **R-29** (2) (June 1980), pp. 126-129.
8. Ferson, S.: *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, Lewis Press, Boca Raton, 2002.
9. Ferson, S.: What Monte Carlo Methods Cannot Do, *Journal of Human and Ecological Risk Assessment* 2 (4) (1996), pp. 990-1007.
10. Ferson, S. and Burgman, M.: Correlations, Dependency Bounds and Extinction Risks, *Biological Conservation* 73 (1995), pp. 101-105.
11. Frank, M.J., Nelsen, R.B., and Schweizer, B.: Best-Possible Bounds for the Distribution of a Sum – a Problem of Kolmogorov, *Probability Theory and Related Fields* 74 (1987), pp. 199-211.
12. Ingram, G.E., Welker, E.L., and Herrmann, C.R.: Designing for Reliability Based on Probabilistic Modeling Using Remote Access Computer Systems, *Proceedings 7th Reliability and Maintainability Conference,* American Society of Mechanical Engineers, 1968, pp. 492-500.

13. Moore, R.: Risk Analysis Without Monte Carlo Methods, Freiburger Intervall-Berichte 84 (1), 1984, pp. 1-48.

14. Nelsen, R.B.: *An Introduction to Copulas,* Lecture Notes in Statistics 139, Springer-Verlag, Heidelberg, 1999.

15. Neumaier, A.: Clouds, Fuzzy Sets, and Probability Intervals, *Reliable Computing,* forthcoming, http://www.mat.univie.ac.at/ ~neum/papers.html.

16. Red-Horse, J. and Benjamin, A.S.: A Probabilistic Approach to Uncertainty Quantification with Limited Information, *Reliability Engineering and System Safety,* forthcoming.

17. Sheblé, G. and Berleant, D.: Bounding the Composite Value at Risk for Energy Service Company Operation with DEnv, an Interval-Based Algorithm, *SIAM Workshop on Validated Computing 2002 Extended Abstracts*, Toronto, May 23-25, pp. 166-171. http://www.public.iastate.edu/~berleant/.

18. Springer, M.D.: *The Algebra of Random Variables,* John Wiley and Sons, New York, 1979.

19. Statool software, http://class.ee.iastate.edu/berleant/home/Research/Pdfs/versions/statool/distribution/index.htm.

20. Tajar, A., Denuit, M., and Lambert, P.: *Copula-Type Representations for Random Couples with Bernoulli Margins*, Discussion Paper 0118, Institut de Statistique, Université Catholique de Luvain, 2001, http://www.stat.ucl.ac.be/ISpublications.html.

21. Williamson, R.C. and Downs, T.: Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds, *International Journal of*

*Approximate Reasoning* 4 (1990), pp. 89-158.

22. Wood, A.J. and Wollenberg, B.F.: *Power Generation, Operations, and Control, 2nd ed.,* Wiley, Hoboken, 1996

# CHAPTER 3. ENVELOPES AROUND CUMULATIVE DISTRIBUTION FUNCTIONS FROM INTERVAL PARAMETERS OF STANDARD CONTINUOUS DISTRIBUTIONS

A paper published in the Proceedings of North American Fuzzy Information Processing Society (NAFIPS 2003), 407-412, 2003.

Jianzhong Zhang and Daniel Berleant

## Abstract

A cumulative distribution function (CDF) states the probability that a sample of a random variable will be no greater than a value x, where x is a real value. Closed form expressions for important CDFs have parameters, such as mean and variance. If these parameters are not point values but rather intervals, sharp or fuzzy, then a single CDF is not specified. Instead, a family of CDFs is specified. Sharp intervals lead to sharp boundaries ("envelopes") around the family, while fuzzy intervals lead to fuzzy boundaries. Algorithms exist [12] that compute the family of CDFs possible for some function g(v) where v is a vector of distributions or bounded families of distribution. We investigate the bounds on families of CDFs implied by interval values for their parameters. These bounds can then be used as inputs to algorithms that manipulate distributions and bounded spaces defining families of distributions (sometimes called probability boxes or p-boxes). For example, problems defining inputs this way may be found in [10,12]. In this paper, we present the bounds for the families of a few common CDFs when parameters to those CDFs are intervals.

## Introduction

Uncertainties are ubiquitous in realistic models. Handling such uncertainty is an important issue in reliable computing. A variety of methods have been developed to deal with this problem [11, 12]. Compared with the traditional method, Monte Carlo, these methods are not subject to noise effects due to randomness that can affect the results obtained from Monte Carlo methods (Ferson 1996 [6]). Such methods offer principled approaches to manipulating uncertain quantities in the presence of 2nd-order uncertainties such as uncertainties in parameters of distributions.

Accurate modeling all too often requires handling the situation that exact distributions are not known, though some information about them is known. To handle this situation, Smith used limited information about distributions to get bounds on the expected value of an arbitrary objective function (1995 [14]). The method is based on moments of distributions. One way to express that information is with interval-valued parameters to standard distributions [10]. Ferson presented some initial results, including examples of envelopes for families of normal distributions defined by interval-valued means and variances, uniform distributions, and Weibull distributions (2003 [7]). The need to formalize and generalize such results helps motivate the present work.

In general, simulation can be adopted to estimate envelopes for distributions with interval parameters. But having CDF envelopes available in closed form can save considerable computation over approximating them when needed using MC simulation. Thus we seek to obtain the left and right envelopes around the family of CDFs for a random variable whose distribution is expressed in closed form with interval parameters.

Then these envelopes can be used to compute envelopes around derived distributions using our Distribution Envelope Determination (DEnv) algorithm or another algorithm [1-5, 8, 12]).

## Deriving envelopes analytically

In order to determine CDF envelopes by analyzing the effect of parameters to the underlying CDF, the core idea is to find the minimum and maximum boundaries, expressed in closed form, for CDFs of random variables when parameter values are specified to be within particular intervals. Then, the curve for the CDF implied by any numerically valued parameters that fall within their respective intervals, will be wholly between those boundaries.

Denote a parameterized CDF with $H(x,\vec{\theta}) = F(x)$ where $x$ is a value of the random variable and $\vec{\theta}$ is a vector of one or more parameters. Assume that each $\theta_i$ is not necessarily specified to be a specific numerical value, but instead can be an interval $\psi_i$. We wish to find the left envelope function $E_l(x) = \max_{\theta_i \in \psi_i \forall i} H(x,\vec{\theta})$ and the right envelope function $E_r(x) = \min_{\theta_i \in \psi_i \forall i} H(x,\vec{\theta})$.

If $H(x,\vec{\theta})$ is monotonic function about each $\theta_i$, the results are derived as follows. Let $\underline{\theta_i}$ be the minimum value of $\psi$, and $\overline{\theta_i}$ be the maximum value of $\psi$. If $H(x,\vec{\theta})$ is non-decreasing, $E_l(x) = H(x,\overline{\theta_1},...,\overline{\theta_I})$ given I parameters, and $E_r(x) = H(x,\underline{\theta_1},...,\underline{\theta_I})$. If $H(\theta)$ is non-increasing, $E_l(x) = H(x,\underline{\theta_1},...,\underline{\theta_I})$ and $E_r(x) = H(x,\overline{\theta_1},...,\overline{\theta_I})$.

If $H(x,\vec{\theta})$ is not monotonic, the solution is to partition the domain into regions within which it is monotonic. Different portions of $E_l$ and $E_r$ may derive from different regions and have different functions. In the next section we discuss envelopes which may be derived without partitioning the domain, and in the subsequent section we discuss envelopes for which partitioning is necessary.

## Envelopes derivable without partitioning

This section gives envelopes for a few common distributions for which the values of the parameters that lead to envelopes whose functions do not depend on the value of the distribution's argument $x$. We first discuss how to get the envelopes for exponential distributions. Then we give the results for uniform and triangular distributions.

### Exponential distribution

The density function of an exponential distribution is

$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$ if $x \geq 0$, parameterized with $\beta > 0$.

From the density function, we can get the cumulative probability function by integrating the density function.

$$F(x) = \int_{-\infty}^{0} f(t)dt = \int_{0}^{x} \frac{1}{\beta} e^{-\frac{t}{\beta}} dt = \int_{0}^{x/\beta} \frac{1}{\beta} e^{-y} d(\beta y)$$
$$= \int_{0}^{x/\beta} e^{-y} dy = -e^{-y} \Big|_{0}^{x/\beta} = -e^{-\frac{x}{\beta}} - (-e^{-0}) = 1 - e^{-\frac{x}{\beta}}$$

if $x \geq 0$.

Next we will show how this parameter affects the probability at given value. Consider the parameterized version of $F(x)$, which is $G(x,\beta)$. $G(x,\beta) = 1 - e^{-\frac{x}{\beta}} = 1 - \frac{1}{e^{x/\beta}}$, $\beta > 0$. It is clear that $G(x,\beta)$ is a decreasing function of $\beta$.

For fixed $x$, if $\beta$ increases, $G(x,\beta)$ will decrease, so we get a bigger probability if we use a smaller value of $\beta$. Assume $\beta$ belongs to interval $[a,b]$. Then $E_l(x) = 1 - e^{-\frac{x}{a}}, x \geq 0$, and $E_r(x) = 1 - e^{-\frac{x}{b}}, x \geq 0$.

For any other $\beta$ in $[a,b]$, the CDF $G(x,\beta)$ must lie between envelopes $E_l(x)$ and $E_r(x)$. The following figure shows the case when a=1 and b=3.

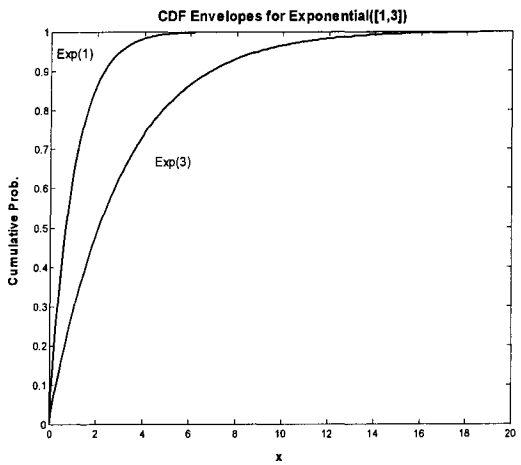



Figure 1. Exponential envelopes $E_l(x)$=Exp(1) and $E_r(x)$=Exp(3) are shown; $\beta \in [1,3]$.

Now consider another parameter, the location parameter. Since decreasing the location parameter would move the CDF to the left, and increasing it would move it to the right, the left envelope function would use the minimum value of the location parameter and the right envelope function would use its maximum value. Thus if both the location parameter and parameter $\beta$ were given as intervals, the left envelope would be derived from the low values of both parameters and the right envelope would be derived from their high values.

## Uniform distribution.

If a random variable $X$ follows the uniform distribution, 2 parameters may be used to describe it: $X_{\min}$ and $X_{\max}$. $X_{\min}$ is the minimum value and $X_{\max}$ is the maximum value possible for samples of $X$. The relationship between these two parameters is $X_{\min} < X_{\max}$ and the density function is

$$f(x) = \frac{1}{X_{\max} - X_{\min}}, \quad X_{\min} \le x \le X_{\max}.$$

From the density function, we can get the cumulative distribution function:

$$F(x) = \frac{x - X_{\min}}{X_{\max} - X_{\min}}, \quad X_{\min} \le x \le X_{\max}.$$

Define a parameterized version of $F(x)$ as

$G(x, X_{\min}, X_{\max})$. Since $G$ decreases as $X_{\min}$ and $X_{\max}$ increase, the smaller the parameters the higher the cumulative probability. In general, if we know 2 intervals $[a,b]$ $[c,d]$ for $X_{\min}$ and $X_{\max}$ respectively, then

$$E_l(x) = \begin{cases} \frac{x-a}{c-a} & c > x \ge a \\ 1 & x \ge c \end{cases} \text{and } E_r(x) = \begin{cases} \frac{x-b}{d-b} & d > x \ge b \\ 1 & x \ge d \end{cases}$$

For any other values of the parameters in those intervals, the CDFs will lie between the envelope CDFs $E_l$ and $E_r$. The following figure depicts the situation when $a$=1, $b$=2, $c$=5, and $d$=6.

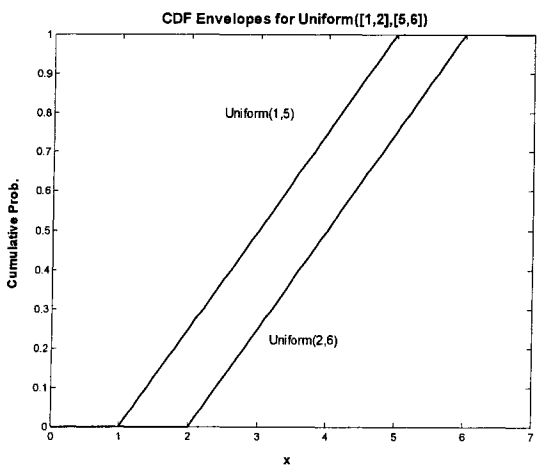



Figure 2. Envelopes based on parameters of the uniform distribution.

## Triangular distribution

Three parameters describe triangular probability density functions. They are $X_{min}$, $X_{mod}$, and $X_{max}$. $X_{min}$ is the minimum value of $X$, $X_{max}$ is the maximum value of $X$, and $X_{mod}$ is the mode value of $X$. The relationship between these values is

$$X_{min} \le X_{mod} \le X_{max} \text{ and } X_{min} < X_{max}.$$

Its density function is

$$f(x)=\frac{2*(x-X_{min})}{(X_{max}-X_{min})(X_{mod}-X_{min})}, X_{min}\le x\le X_{mod}$$

$$f(x)=\frac{2*(X_{max}-x)}{(X_{max}-X_{min})(X_{max}-X_{mod})}, X_{mod}< x\le X_{max}$$

From the density function, we can derivate its cumulative probability function.

$$F(x)=\frac{(x-X_{min})^2}{(X_{max}-X_{min})(X_{mod}-X_{min})}, X_{min}\le x\le X_{mod}$$

$$F(x)=1-\frac{(X_{max}-x)^2}{(X_{max}-X_{min})(X_{max}-X_{mod})}, X_{mod}< x\le X_{max}$$

Based on these CDFs, we can conclude that the smaller the parameter, the higher the cumulative probability $F$. Let us describe the parameters with three intervals $[a,b]$, $[c,d]$, and $[e,f]$ for $X_{min}$, $X_{mod}$ and $X_{max}$ respectively, where $a<b<c<d<e<f$. then $E_l(x)$ and $E_r(x)$ can be written as follows.

$$E_l(x)=\begin{cases}\dfrac{(x-a)^2}{(e-a)(c-a)} & a\le x\le c\\ 1-\dfrac{(e-x)^2}{(e-a)(e-c)} & c< x\le e\\ 1 & x>e\end{cases}$$

and

$$E_r(x) = \begin{cases} \dfrac{(x-b)^2}{(f-b)(d-b)} & b \le x \le d \\ 1 - \dfrac{(f-x)^2}{(f-b)(f-d)} & d < x \le f \\ 1 & x > f \end{cases}$$

The space between this pair of envelopes will contain all other CDFs generating from parameters within those intervals. The following figure demonstrates this situation for $a$=1, $b$=2, $c$=3, $d$=4, $e$=5, and $f$=6.

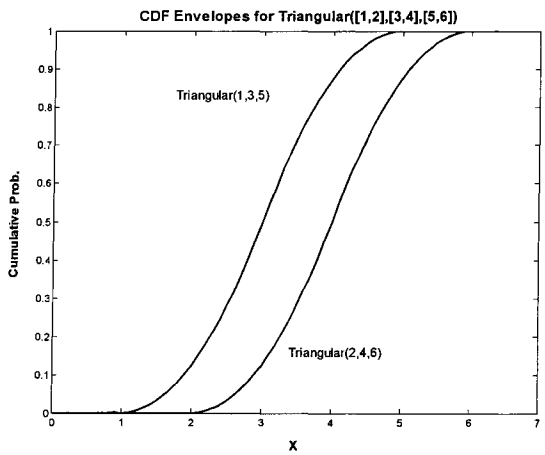



Figure 3. Envelopes around the CDFs of triangular density functions, derived from interval constraints on its parameters.

## Envelopes requiring partitioning to derive

In this section, we present envelopes for the Cauchy, normal, and lognormal distributions.

### Cauchy distribution

Let us use two parameters to describe the Cauchy distribution, a location parameter $\mu$, and a scale parameter $\sigma$. Here $\mu \in R$ and $\sigma > 0$.

The density function of Cauchy distribution is

$$f(x)=\frac{1}{\pi}\frac{\sigma}{\sigma^2+(x-\mu)^2},\ x\in R$$

From the density function, we can get its cumulative probability function by integrating its density function.

$$F(x)=\int_{-\infty}^{x}f(t)dt=\int_{-\infty}^{x}\frac{1}{\pi}\frac{\sigma}{\sigma^2+(t-\mu)^2}dt=\int_{-\infty}^{x}\frac{1}{\pi}\frac{\sigma}{\sigma^2(1+(\frac{x-\mu}{\sigma})^2)}dt$$

$$=\int_{-\infty}^{x}\frac{1}{\pi}\frac{1}{\sigma(1+(\frac{x-\mu}{\sigma})^2)}dy=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\pi\sigma(1+y^2)}d(\mu+\sigma y)=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\pi}\frac{1}{(1+y^2)}dy$$

$$=\frac{1}{\pi}\tan^{-1}y\Big|_{-\infty}^{\frac{x-\mu}{\sigma}}=\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}-\frac{1}{\pi}\tan^{-1}(-\infty)$$

$$=\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}-\frac{1}{\pi}(-\frac{\pi}{2})$$

$$=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$$

Let $y=\frac{x-\mu}{\sigma}$ and consider the resulting function $G(y)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}y$. Let us consider the interval for each parameter in turn.

**Location parameter $\mu$**

$H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is a decreasing function of $\mu$ since it is given that $\sigma>0$. Hence the smaller the value of $\mu$, the higher the value of $H$ and hence the higher the cumulative probability for a given value of $x$.

**Scale parameter $\sigma$**

The effect on $y=\frac{x-\mu}{\sigma}$ of changing $\sigma$ depends on the sign of $x-\mu$. If $x-\mu>0$, then $y$ decreases as $\sigma$ increases, so $G(y)$ also decreases. So $H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is a decreasing

function of $\sigma$ for $x-\mu>0$. If $x-\mu<0$, then increasing $\sigma$ increases $y$, so $G(y)$ also increases. So $H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is an increasing function of $\sigma$ for $x-\mu<0$.

Combining the two situations just noted, we have to use different formulas for different regions of an envelope, with the regions meeting at $x=\mu$. Consider intervals $[a,b]$ and $[c,d]$ for $\mu$ and $\sigma$ respectively. Then we get the following envelope functions.

$$E_l(x)=\begin{cases}\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-a}{c} & x\geq a\\ \frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-a}{d} & x<a\end{cases}$$

$$E_r(x)=\begin{cases}\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-b}{d} & x\geq b\\ \frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-b}{c} & x<b\end{cases}$$

For any other values of the parameters consistent with their intervals, the CDF must lie between the region enclosed by the two envelope CDFs. When $a=-5$, $b=5$, $c=9$ and $d=25$, the following figure shows the envelopes.

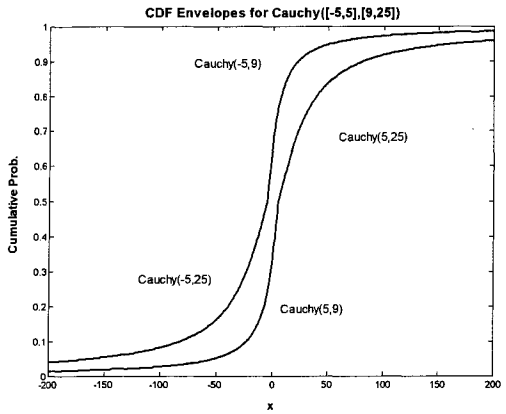



Figure 4. Envelopes around the Cauchy distribution implied by intervals for its two parameters. Each envelope function has two regions which meet at a non-differentiable point, $x=a$ for $E_l$ and $x=b$ for $E_r$.

## Normal distribution

There are two parameters sufficient to describe the normal distribution, the location parameter $\mu$ and the scale parameter $\sigma$. Possible values for these parameters are $\mu \in R$ and $\sigma > 0$.

The density function of the normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } x \in R.$$

From the density function, we characterize the cumulative function as follows.

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)dt$$

Define $y = \frac{t-\mu}{\sigma}$. Then

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{y=-\infty}^{y=\frac{x-\mu}{\sigma}} \exp\left(\frac{-y^2}{2}\right)d(\sigma y + \mu) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \exp\left(\frac{-y^2}{2}\right)\sigma dy$$

$$= \frac{\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \exp\left(\frac{-y^2}{2}\right)dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{\frac{-y^2}{2}} dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{\frac{-y^2}{2}} dy = H(w)$$

where $w = \frac{x-\mu}{\sigma}$. $H(w)$ is an increasing function of $w$ since $e$ to any power is positive. So by considering the direction of change in $w$ caused by changing $\mu$ or $\sigma$, we can conclude $F(x)$ changes in the same direction.

For $w$, and so for $H(w)$, the smaller $\mu$ is the bigger $w$ and $H$ are. The smaller $\sigma$ (and therefore $\sigma^2$ since $\sigma$ is positive) is, the bigger $w$ and $H$ are for $x > \mu$, and the smaller $w$ and $H$ are for $x < \mu$.

In general, consider 2 intervals $[a,b]$, $[c,d]$ for $\mu$ and $\sigma^2$ respectively. $E_l(x)$ and $E_r(x)$ are

$$E_l(x) = \begin{cases} Normal\ (a,c) & x \geq a \\ Normal\ (a,d) & x < a \end{cases}$$

and

$$E_r(x) = \begin{cases} Normal\ (b,d) & x \geq b \\ Normal\ (b,c) & x < b \end{cases}$$

where $Normal(\mu,\sigma^2)$ is the CDF of the normal distribution with mean $\mu$ and variance $\sigma^2$.

For any other values of the parameters in their intervals, the CDF must be within the region enclosed by the two envelope CDFs $E_l$ and $E_r$. The figure below shows the CDF envelopes for $a$=1, $b$=2, $c$=9 and $d$=25.

## Lognormal distribution

We parameterize the lognormal distribution as in Siegrist (2002 [13]), one of several alternatives [9]. This parameterization has two parameters, $\mu$ and $\sigma$. Here $\mu \in R$ and $\sigma > 0$. The density function of the lognormal distribution then is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Let $z = \ln x$. Then $z$ is normally distribution. Thus we can apply the results from the case of the normal distribution here. Consequently for $z$, the smaller the value of $\mu$ the higher the cumulative probability, and the lower $\sigma$ the higher the cumulative probability is if $z \geq \mu$ and the lower the cumulative probability is if $z < \mu$. To derive results for the original argument $x$ from these inequalities for $z$=ln $x$, the term ln $x$ may be substituted for $z$ and the inequalities solved for $x$.

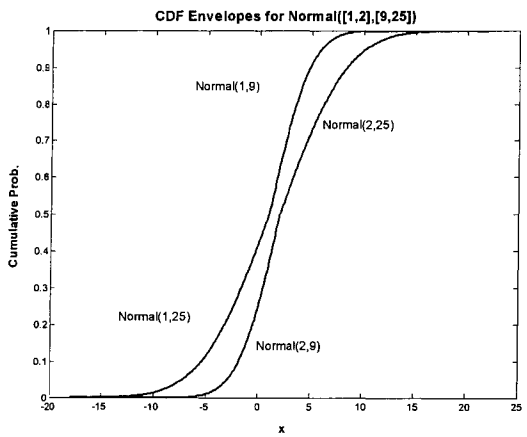



Figure 5. Envelopes around the normal distribution implied by intervals for its location and scale parameters. Each envelope function has two regions which meet at a non-differentiable point, $x=a$ for $E_l$ and $x=b$ for $E_r$.

Applying those steps yields the following formulation. The smaller $\mu$ is, the higher the cumulative probability. The smaller $\sigma$ is, the higher the cumulative probability is if $x \geq e^{\mu}$ and lower the cumulative probability is if $x < e^{\mu}$. The same rules apply to $\sigma^2$ as for $\sigma$ since $\sigma > 0$.

We can now specify intervals for $\mu$ and $\sigma^2$, the endpoints of which can be used to state the equations of the envelopes $E_l$ and $E_r$. Let $\mu$ and $\sigma^2$ be values in $[a,b]$ and $[c,d]$ respectively. Then

$$E_l(x) = \begin{cases} LN & (a,c) & x \geq e^a \\ LN & (a,d) & x < e^a \end{cases}$$

and

$$E_r(x) = \begin{cases} LN & (b,d) & x \geq e^b \\ LN & (b,c) & x < e^b \end{cases}$$

where $LN(\mu,\sigma^2)$ is the CDF of the lognormal distribution with parameters $\mu$ and $\sigma$. As an example, let $a$=3, $b$=4, $c$=0.1, and $d$=0.3. Then the envelopes are shown in the following figure.

## Discussion: fuzzy interval parameters

The results given may be generalized to the case of parameters described with fuzzy intervals. If one parameter is a fuzzy interval, then each cut set of that interval yields a pair of envelopes. A nested series of envelopes results. A vertical slice through the graph then yields a fuzzy interval for the cumulative probability at a given value on the horizontal axis. A horizontal slice through the graph yields a fuzzy interval for the value on the horizontal axis for which the cumulative probability is a particular value.

## Conclusion

We analytically derive envelopes for a variety of standard distributions with interval-valued parameters. For some distributions the envelopes have a non-differentiable point. For

other distributions, we have not yet been able to derive envelopes analytically. Since there are important distributions which are among those we have not discussed, further work is needed in this direction.

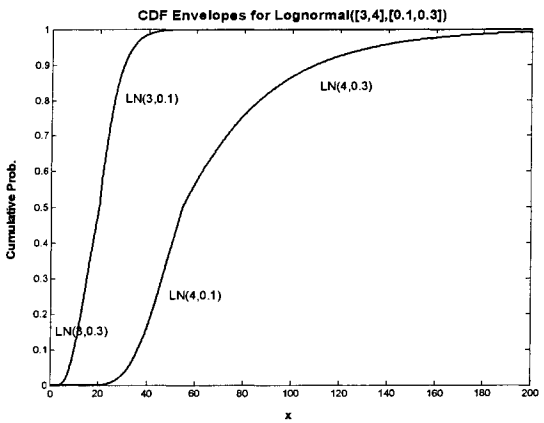



Figure 6. Envelopes around the lognormal distribution implied by intervals for its $\mu$ and $\sigma$ parameters. Values given are for $\mu$ and $\sigma^2$. Each envelope function has two regions which meet at a non-differentiable point, $x=e^a$ for $E_l$ and $x=e^b$ for $E_r$.

## References

[1] Berleant, D. Automatically verified reasoning with both intervals and probability density functions. *Interval Computations* (1993 No. 2), pp. 48-70, http://www.public.iastate.edu/~berleant/.

[2] Berleant, D. and C. Goodman-Strauss. Bounding the results of arithmetic operations on random variables of unknown dependency using intervals. *Reliable Computing* 4(2) (1998), pp. 147-165, http://www.public.iastate.edu/~berleant/.

[3] Berleant, D, L. Xie, and J. Zhang. Statool: a tool for distribution envelope determination (DEnv), an interval-based algorithm for arithmetic on random variables. *Reliable Computing* 9 (2) (2003), pp. 91-108, http://www.public.iastate.edu/~berleant/.

[4] Berleant, D and J. Zhang. Representation and problem solving with the Distribution Envelope Determination (DEnv) method. *Reliability Engineering and System Safety*, **85** (1-3) (2004), pp. 153-168, http://www.public.iastate.edu/~berleant/.

[5] Berleant, D and J. Zhang. Using Pearson correlation to improve envelopes around the distributions of functions. *Reliable Computing*, **10** (2) (2004), pp. 139-161, http://www.public.iastate.edu/~berleant/.

[6] Ferson, S. What Monte Carlo methods cannot do. *Journal of Human and Ecological Risk Assessment* 2 (4)(1996), pp. 990-1007

[7] Ferson, S., V. Kreinovich, L. Ginzburg., D. Myers, and K. Sentz. Constructing Probability Boxes and Dempster-Shafer Structures. *SAND REPORT SAND2002-4015*, Sandia National Laboratories, Jan. 2003.

[8] Neumaier, A. Clouds, fuzzy sets and probability intervals, Reliable Computing **10** (2004), 249-272, http://www.mat.univie.ac.at/~neum/papers.html. See also, On the structure of clouds, submitted, same URL.

[9] *NIST/SEMATECH e-Handbook of Statistical Methods.* Web site http://www.itl.nist.gov/div898/handbook/, as of 2003. Paper at http://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm.

[10] Oberkampf, W., J. Helton, C. Joslyn, S. Wojtkiewicz, and S. Ferson. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety*, **85** (1-3) (2004).

[11] Regan, H., S. Ferson S, and D. Berleant. Equivalence of five methods for bounding uncertainty. *Journal of Approximate Reasoning*, **36** (2004), pp. 1-30.

[12] Sandia National Laboratory. Epistemic Uncertainty Workshop. August 6-7, 2002, Albuquerque, www.sandia.gov/epistemic/eup_workshop1.htm.

[13] Siegrist, K. Virtual Laboratories in Probability and Statistics. Web site http://www.math.uah.edu/statold/, URL http://www.math.uah.edu/statold/special/special14.html.

[14] Smith, J.E. Generalized Chebychev inequalities: theory and application in decision analysis. *Operations Research* (1995) 43: 807-825.

# CHAPTER 4: ARITHMETIC ON RANDOM VARIABLES: SQUEEZING THE ENVELOPES WITH NEW JOINT DISTRIBUTION CONSTRAINTS

A paper published in the Proceedings of the International Symposium on Imprecise Probabilities and Their Applications (ISIPTA '05), 416-422, 2005.

Jianzhong Zhang and Daniel Berleant

## Abstract

Uncertainty is a key issue in decision analysis and other kinds of applications. Researchers have developed a numbers of approaches to address computations on uncertain quantities. When doing arithmetic operations on random variables, an important question has to be considered: the dependency relationships among the variables. In practice, we often have partial information about the dependency relationship between two random variables. This information may result from experience or system requirements. We can use this information to improve bounds on the cumulative distributions of random variables derived from the marginals whose dependency is partially known.

**Keywords.** Uncertainty, arithmetic on random variables, distribution envelope determination (DEnv), joint distribution, dependency relationship, copulas, probability boxes, linear programming, partial information.

## Introduction

Uncertainty is a key issue in decision analysis and other kinds of reasoning. Researchers have developed a numbers of approaches to address computations on uncertain distributions. Some of these approaches are confidence limits (Kolmogoroff 1941), discrete convolutions (i.e. Cartesian products, Ingram 1968), probabilistic arithmetic (Williamson and

Downs 1990), Monte Carlo simulation (Ferson 1996), copulas (Nelsen 1999), stochastic dominance (Levy 1999), clouds (Neumaier 2004), and Distribution Envelope Determination (Berleant and Zhang 2004a).

Belief and plausibility curves, upper and lower previsions, left and right envelopes, and probability boxes designate an important type of representation for bounded uncertainty about distributions. When doing arithmetic operations on random variables that can result in such CDF bounds, an important question has to be considered: the dependency relationships among the variables. Couso et al. (1999) and Fetz and Oberguggenberger (2004) addressed different concepts of independence and their effects on CDF bounds. The copula-based approach can represent many interesting constraints on joint distributions that affect CDF bounds (e.g. Clemen 1999, Embrechts et al. 2003, Ferson and Burgman 1995). The Distribution Envelope Determination (DEnv) method can use Pearson correlation between marginals $X$ and $Y$ to squeeze CDF bounds of random variables derived from these marginals (Berleant and Zhang 2004b). This paper explores some additional constraints on dependency.

In practice, we may have partial information about the dependency relationship between two random variables. This information may result from empirical experience or system requirements. We can use this information to affect bounds on the cumulative distributions of new random variables derived from those whose dependency is partially known.

We focus on the following kinds of partial information.

1. Knowledge about probabilities of specified areas of the joint distribution of the marginals.
2. Knowledge about probabilities of specified ranges of values of the derived random variable.

3. Known relationships (>, <, =) among the probabilities of different areas of the joint distribution of the marginals.
4. Known relationships (>, <, =) among the probabilities of different ranges of the derived random variable.

Our method uses the DEnv algorithm (Berleant and Zhang 2004c).

## Review of the Distribution Envelope Determination (DEnv) algorithm

In this section, DEnv is reviewed briefly and abstractly, following Berleant and Zhang (2004a).

Suppose we have two samples $x$ and $y$ of random variables $X$ and $Y$ with probability density functions $f_x(.)$ and $f_y(.)$. Given a function $g$, a sample $z=g(x,y)$ of random variable $Z$ is derived from $x$ and $y$. DEnv is used to get the distribution of the derived variable $Z$. First, the input PDFs $f_x(.)$ and $f_y(.)$ are discretized by partitioning the support (i.e. the domain over which a PDF is non-zero) of each, yielding intervals $\mathbf{x}_i$, $i$=1…$m$, and $\mathbf{y}_j$, $j$=1…$n$. Each $\mathbf{x}_i$ is assigned a probability

$p_{\mathbf{x}_i} = p(x \in \mathbf{x}_i) = \int_{x_0=\underline{\mathbf{x}_i}}^{\overline{\mathbf{x}_i}} f_x(x_0)dx_0$, where interval-valued symbols are shown in bold, and interval $\mathbf{x}_i$ has lower bound $\underline{\mathbf{x}_i}$ and upper bound $\overline{\mathbf{x}_i}$. Similarly, each $\mathbf{y}_j$ is assigned a probability

$p_{\mathbf{y}_j} = p(y \in \mathbf{y}_j) = \int_{y_0=\underline{\mathbf{y}_j}}^{\overline{\mathbf{y}_j}} f_y(y_0)dy_0$. The $\mathbf{x}_i$'s and $\mathbf{y}_j$'s and their probabilities form the marginals of a discretized joint distribution called a joint distribution tableau (Table 1), the interior cells of which each contain two items. One is a probability mass

$p_{ij} = p(x \in \mathbf{x}_i \wedge y \in \mathbf{y}_j)$. If $x$ and $y$ are independent then $p_{ij} = p(x \in \mathbf{x}_i) \cdot p(y \in \mathbf{y}_j) = p_{\mathbf{x}_i} \cdot p_{\mathbf{y}_j}$, where $p_{\mathbf{x}_i}$ is defined as $p(x \in \mathbf{x}_i)$ and $p_{\mathbf{y}_j}$ as

$p(y \in \mathbf{y}_j)$. The second item is an interval that bounds the values $z=g(x,y)$ may have, given that $x \in \mathbf{x}_i \wedge y \in \mathbf{y}_j$. In other words, $\mathbf{z}_{ij}=\mathbf{g}(\mathbf{x}_i, \mathbf{y}_j)$.

| $y\downarrow$ $x\rightarrow$ <br> $z=g(x,y)\ \searrow$ | ... | $\mathbf{x}_i$ <br> $p_{\mathbf{x}_i} = p(x \in \mathbf{x}_i)$ | ... |
|---|---|---|---|
| ... | ... | ... | ... |
| $\mathbf{y}_j$ <br> $p_{\mathbf{y}_{ij}} = p(y \in \mathbf{y}_j)$ | ... | $\mathbf{z}_{ij}=\mathbf{g}(\mathbf{x}_i,\mathbf{y}_j)$ <br> $p_{ij} = p(x \in \mathbf{x}_i \wedge y \in \mathbf{y}_j)$ | ... |
| ... | ... | ... | ... |

Table 1: General form of a joint distribution tableau for random variables $X$ and $Y$.

To better characterize the CDF $F_z(.)$, we next convert the set of interior cells of the joint distribution tableau into cumulative form. Because the distribution of each probability mass $p_{ij}$ over its interval $\mathbf{z}_{ij}$ is not defined by the tableau, values of $F_z(.)$ cannot be computed precisely. However they can be bounded. DEnv does this by computing the analogous interval-valued function $\mathbf{F}_z(.)$ as

$$\underline{\mathbf{F}_z}(z_0) = \sum_{i,j|\mathbf{z}_{ij} \leq z_0} p_{ij} \quad \text{and} \quad \overline{\mathbf{F}_z}(z_0) = \sum_{i,j|\underline{\mathbf{z}_{ij}} \leq z_0} p_{ij}, \qquad (1)$$

resulting in right and left envelopes respectively bounding $F_z(.)$.

An additional complication occurs if the dependency relationship between $x$ and $y$ is unknown. Then the values of the $p_{ij}$'s are underdetermined, so equations (1) cannot be evaluated. However, the $p_{ij}$'s in column $i$ of a joint distribution tableau must sum to $p_{\mathbf{x}_i}$ and the $p_{ij}$'s in row $j$ must sum to $p_{\mathbf{y}_j}$ giving three sets of constraints: $p_{\mathbf{x}_i} = \sum_j p_{ij}$, $p_{\mathbf{y}_j} = \sum_i p_{ij}$, and $p_{ij} \geq 0$, for $i=1\ldots m$, $j=1\ldots n$. These constraints are all linear, and so may be optimized by

linear programming. Linear programming takes as input linear constraints on variables, which in this case are the $p_{ij}$'s, and an expression in those variables to minimize, for example, $\underline{\mathbf{F}}_z(z_0) = \sum_{i,j|\mathbf{z}_{ij} \le z_0} p_{ij}$ in equations (1) for some given value $z_0$. The output produced is the minimum value possible for $\underline{\mathbf{F}}_z(z_0)$, such that the values assigned to the $p_{ij}$'s are consistent with the constraints. $\overline{\mathbf{F}}_z(z_0) = \sum_{i,j|\mathbf{z}_{ij} \le z_0} p_{ij}$ in equations (1) is maximized similarly. These envelopes are less restrictive (i.e. are farther apart) than when the $p_{ij}$'s are fully determined by an assumption of independence or some other given dependency relationship (in which case linear programming would not be needed).

These ideas could be generalized to $n$ marginals, which would require an $n$-dimensional joint distribution tableau.

Next, we examine additional constraints that can be used to try to squeeze the envelopes closer together.

## Knowledge about probabilities over specified areas of the joint distribution

Suppose we have information about the probabilities over given portions of the joint distribution. It could be that we know the probabilities exactly or perhaps we only know bounds on these values.

This problem breaks down into two major situations:

- ***Single-cell constraints***, where the probability of one $p_{ij}$ is known in a joint distribution tableau, section 3.1.
- ***Multiple-cell constraints***, where our knowledge about probability spans more than one $p_{ij}$, section 3.2.

For ***multiple-cell constraints***, there are two subcategories:

- ***Area specified***, where we have knowledge about a sum $p_{ij} + \ldots + p_{mn}$, section 3.2.1.

- ***Probability of a function of the marginals specified over part of its domain***, where we have knowledge abut the probability of $g(x,y)$ over some interval $k_1 \le g(x,y) \le k_2$, section 3.2.2.

We explore these situations in the following examples. Assume that the marginal distributions of $X$ and $Y$ are known, and define $Z=X+Y$ as in Table 2.

| $x \rightarrow$<br>$y \downarrow \quad z=x+y$ | $\mathbf{x}_1=[x_{1l}, x_{1h}]$<br>$p_{\mathbf{x}_1}$ | ... | $\mathbf{x}_m=[x_{ml}, x_{mh}]$<br>$p_{\mathbf{x}_n}$ |
|---|---|---|---|
| $\mathbf{y}_1=[y_{1l}, y_{1h}]$<br>$p_{\mathbf{y}_1}$ | $\mathbf{z}_{11}=[x_{1l}+y_{1l},$<br>$x_{1h}+y_{1h}]$<br>$p_{11}$ | ... | $\mathbf{z}_{1m}=[x_{ml}+y_{1l},$<br>$x_{mh}+y_{1h}]$<br>$p_{1m}$ |
| ... | ... | ... | ... |
| $\mathbf{y}_n=[y_{nl}, y_{nh}]$<br>$p_{\mathbf{y}_m}$ | $\mathbf{z}_{1n}=[x_{1l}+y_{nl},$<br>$x_{1h}+y_{nh}]$<br>$p_{n1}$ | ... | $\mathbf{z}_{mn}=[x_{ml}+y_{nl},$<br>$x_{mh}+y_{nh}]$<br>$p_{mn}$ |

Table 2: Joint distribution tableau for the marginals $X$ and $Y$, where $Z=X+Y$. Interval $\mathbf{x}_1$ has low bound $x_{1l}$ and high bound $x_{1h}$, and similarly for other intervals.

Note the following row constraints:

$$\sum_{i=1}^{m} p_{ij} = p_{y_j} \text{ for } j=1 \text{ to } n, \qquad (2)$$

and the following column constraints:

$$\sum_{j=1}^{n} p_{ij} = p_{x_i} \text{ for } i=1 \text{ to } m. \qquad (3)$$

These are due to the properties of joint distributions.

The $p_{ij}$'s, $i$=1 to $m$, $j$=1 to $n$, are unknown. However, the row and column constraints limit the freedom of the $p_{ij}$'s significantly. This fact limits the space of feasible solutions for the linear programming problems in the DEnv algorithm. If we can get additional constraints, this space may be limited even more. That means that we could get bigger values for the minimization questions and/or smaller values for the maximization questions than we otherwise would obtain. Recall that in DEnv, the minimization values provide the right envelope and the maximization values provide the left envelope. If minimization outcomes become bigger or maximization outcomes become smaller, the left and right envelopes will become closer to each other. Thus we will get a more tightly specified space of possible CDFs for random variable $Z$, where $Z$ is a function of the marginals. Based on the example of Table 2, we demonstrate the use of constraints resulting from (1) ***single-cell constraints***, (2) ***multiple-cell constraints with area specified***, and (3) ***probability of a function of the marginals specified over part of its domain***, in the following subsections.

## Single-cell constraints

Consider internal cells $\mathbf{z}_{ij}$ (Table 3). If only the row and column constraints hold, the probability $p_{ij}$ of a given cell $\mathbf{z}_{ij}$ is not fully specified, but only constrained to some degree. Let us specify an additional stronger constraint on some $p_{ij}$, that it has some value $p_{ij}=c_{ij}$. This new constraint can be combined with the row and column constraints. This will tend to squeeze envelopes of $Z$ closer together due to the general observation that more constraints tend to produce stronger conclusions.

This situation is relatively strict. To weaken it, the user may specify an inequality for $p_{ij}$ such as $p_{ij} < c_{ij}$ or $\underline{\mathbf{c}_{ij}} \le p_{ij} \le \overline{\mathbf{c}_{ij}}$.

Here is an example. Consider two random variables $X$ and $Y$ having the discretized distribution shown in the joint distribution tableau of Table 3. $Z=X+Y$ is the derived random variable.

| $x\rightarrow$<br>$y\downarrow \ z=x+y$ | $\mathbf{x}_1=[0,1]$<br>$p_{x_1}=0.2$ | $\mathbf{x}_2=[3,4]$<br>$p_{x_2}=0.4$ | $\mathbf{x}_3=[5,6]$<br>$p_{x3}=0.4$ |
|---|---|---|---|
| $\mathbf{y}_1=[0,1]$<br>$p_{y_1}=0.4$ | $\mathbf{z}_{11}=[0,2]$<br>$p_{11}$ | $\mathbf{z}_{12}=[3,5]$<br>$p_{12}$ | $\mathbf{z}_{13}=[5,7]$<br>$p_{13}$ |
| $\mathbf{y}_2=[3,4]$<br>$p_{y_2}=0.6$ | $\mathbf{z}_{21}=[3,5]$<br>$p_{21}$ | $\mathbf{z}_{22}=[6,8]$<br>$p_{22}$ | $\mathbf{z}_{23}=[8,10]$<br>$p_{23}$ |

Table 3: A joint distribution tableau for $Z=X+Y$.

Figures 1 & 2 show the CDFs of marginals $X$ and $Y$ implied by Table 3.

Suppose $p_{11}= 0.16$ is given (a single-cell constraint). If it is included with the original set of row and column constraints, the envelopes will tend to be squeezed together.

The sum of $X$ and $Y$ *without* the single-cell constraint $p_{11}=0.16$ is shown in Figure 3, while the sum *with* the constraint $p_{11}=0.16$ is shown in Figure 4. It is clear that the envelopes for $Z=X+Y$ are significantly narrowed as a result of this new constraint. If a weaker single-cell constraint is substituted for $p_{11}=0.16$, the envelopes are likely to be narrower than those of Figure 3, but wider than those of Figure 4. For example, Figure 5 shows the envelopes resulting from the constraint $0.15 \leq p_{11} \leq 0.17$.

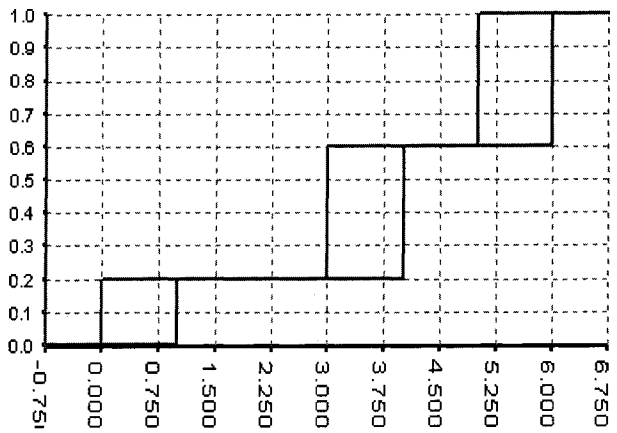



Figure 1. CDF envelopes for X.

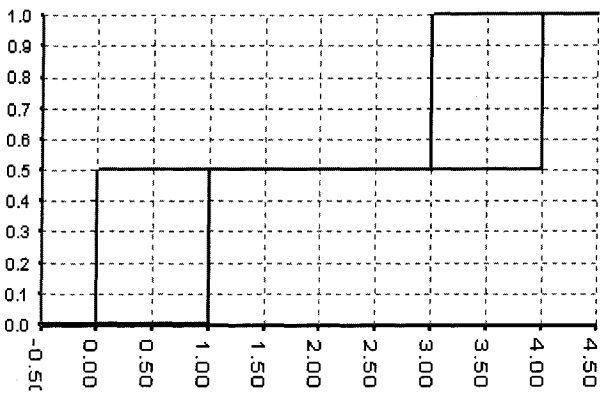



Figure 2. CDF envelopes for Y.

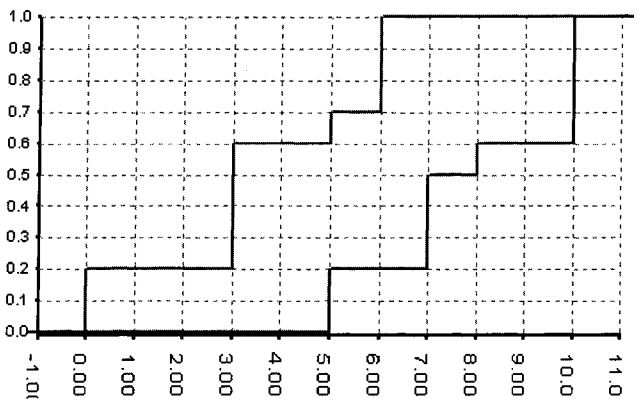



Figure 3. $\mathbf{F}_Z(.)$ for $Z=X+Y$ without any extra constraints.

The envelopes shown in Figure 5 are closer together than those in Figure 3, but further apart than those in Figure 4.

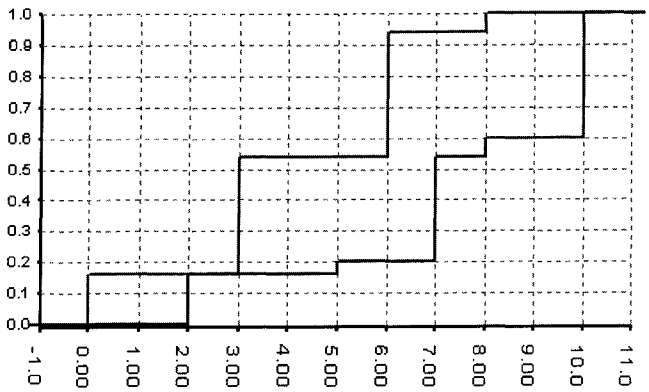



Figure 4. $\mathbf{F}_z(.)$ for $Z=X+Y$ with the single-cell constraint $p_{11}=0.16$.

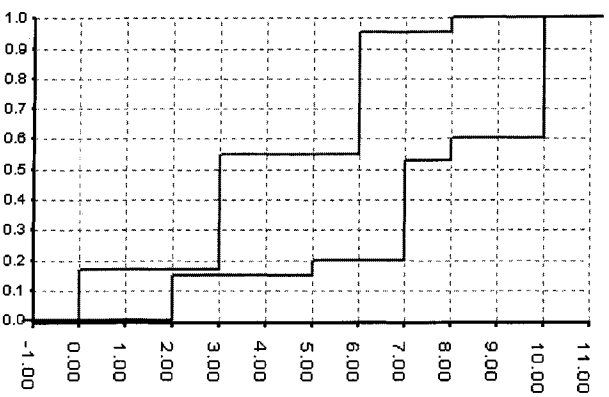



Figure 5. $\mathbf{F}_z(.)$ for $Z=X+Y$ with the single-cell constraint $0.15 \leq p_{11} \leq 0.17$.

## Multiple-cell constraints

In section 3.1 we examined the situation where extra probabilistic information is available for *one* cell. This section explains the situation when extra probabilistic information is connected with a *set* of cells. This generalizes the case of the single-cell constraint. This situation includes two kinds of constraints: we will call these the ***area specified*** constraint

and the ***probability of a function of the marginals specified over part of its domain*** constraint.

### Area specified constraint

Here, a known probability describes the sum of the probabilities of multiple $p_{ij}$'s in the joint distribution tableau, instead of just one $p_{ij}$. This could occur if the probability of a certain region of the joint distribution is given, and that region spans multiple cells of the joint distribution tableau. However, the idea of constraining the probability of a summed probability of a number of cells generalizes to any set of cells, not just ones representing a contiguous region of the joint distribution.

For example, suppose $p_{11}+p_{12}+p_{21}=0.5$ in Table 3. Figure 6 shows the result of including this constraint with the row and column constraints of that table.

Compared with Figure 3, which has no extra constraints, this result has narrower envelopes.

### Probability of a function of the marginals specified over part of its domain

Instead of focusing on the probability of areas of the joint distribution, as with the ***area specified*** constraint, this constraint focuses on probabilities of ranges of $Z$, where $z=g(x,y)$. To illustrate this situation, suppose that $p_z = p(z \in [0,5]) = 0.5$, where $z$ is a sample of $Z$ and $Z=X+Y$. The joint distribution tableau is as in Table 3. Then $p_z$ *must* include $p_{11}$, $p_{12}$, and $p_{21}$ because $\mathbf{z}_{11} = [0,2] \subset [0,5]$, $\mathbf{z}_{12} = [3,5] \subset [0,5]$, and $\mathbf{z}_{21} = [3,5] \subset [0,5]$. For all other $\mathbf{z}_{ij}$, $\mathbf{z}_{ij} \not\subset [0,5]$, so the probability of each such $\mathbf{z}_{ij}$ possibly could be distributed outside of [0,5], hence those $\mathbf{z}_{ij}$ might not contribute to $p_z$. Thus we have that $p_z = 0.5$ and $p_z \geq p_{11}+p_{12}+p_{21}$. This gives the constraint $0.5 \geq p_{11}+p_{12}+p_{21}$.

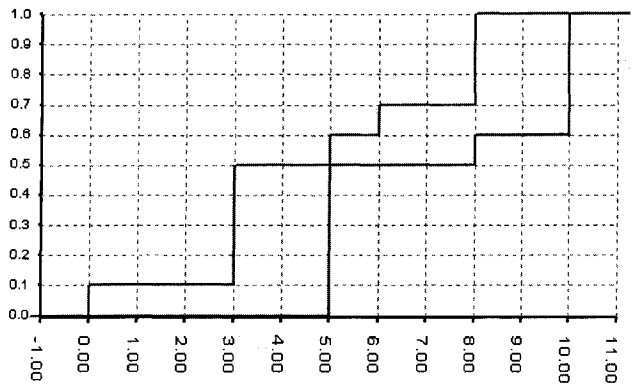



Figure 6. Results for $\mathbf{F}_z(.)$ using the area specified constraint of $p_{11}+p_{12}+p_{21}=0.5$.

Similarly, $p_z$ *might* also include $p_{13}$. This would occur if $\mathbf{z}_{13}$ has its probability distributed as an impulse at its low bound of 5. This gives $0.5 \leq p_{11}+p_{12}+p_{21}+p_{13}$. These two constraints, $p_{11}+p_{12}+p_{21} \leq 0.5$ and $p_{11}+p_{12}+p_{21}+p_{13} \geq 0.5$, result from the constraint $p_z = p(z \in [0,5]) = 0.5$. Figure 7 shows the results using these constraints.

The envelopes in Figure 7 are narrower than in Figure 3, due to the effects of the constraint that $p_z = p(z \in [0,5]) = 0.5$.

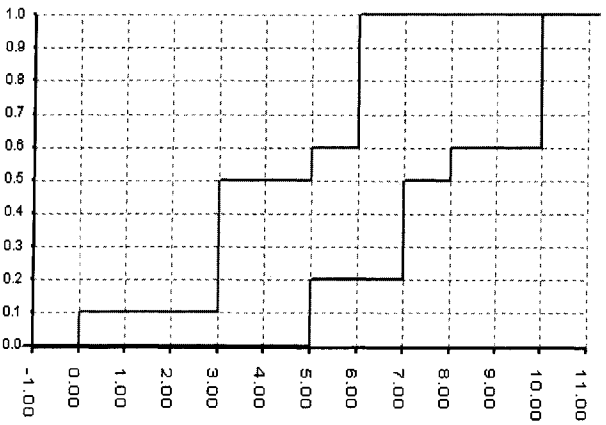



Figure 7. If $p_z = p(z \in [0,5]) = 0.5$, these envelopes result for $\mathbf{F}_z(.)$.

## Known relationship among different areas of the joint distribution constraints

In the previous section we showed how probabilities of certain areas of a joint distribution can be used to narrow envelopes. In this section, we show how *relationships* among probabilities of different areas of the joint distribution can also be used to improve the CDF envelopes.

### Unimodality constraint

If we know that the joint distribution is unimodal, this implies a set of relationships among different areas. For example, the fact that the probability density at the mode point is higher than it is in other areas implies constraints on the $p_{ij}$'s of Table 3. Define random variable $Z$ as the sum of $X$ and $Y$ as in Table 2. The row and column constraints are in equations (2) & (3).

If we also know that $X$ and $Y$ have a unimodal joint distribution and that the mode point is in cell $kl$, the probability $p_{kl}$ will be the result of a higher probability density than the other $p_{ij}$'s. Mathematically, $p_{kl} \geq p_{ij}$, $i \neq k$ and/or $j \neq l$, assuming the intervals $\mathbf{z}_{ij}$ have equal widths and do not overlap. If they do not have equal widths and/or they overlap, similar statements can be made that correct for the differences in widths and that take overlaps into account.

Now we have a set of new constraints. These constraints tend to decrease the area of the feasible solutions, narrowing the CDF envelopes.

Consider Table 3 again. If there is information about which cell $\mathbf{z}_{ij}$ contains the mode point, extra constraints may be derived. Suppose the mode point is in $\mathbf{z}_{23}$. Then the probability of $p_{23}$ is greater than that of any other $p_{ij}$. Thus, $p_{23} \geq p_{ij}$, $i \neq 2$ or $j \neq 3$.

These constraints decrease the feasible solution range of original problem, enabling better envelopes to be obtained. Here are all the constraints including the new ones:

$$\sum_{i=1}^{2} p_{ij} = p_{y_j} \quad \text{for } j=1 \text{ to } 3,$$

$\sum_{j=1}^{3} p_{ij} = p_{x_i}$ for $i$=1 to 2,

$p_{23} \geq p_{ij}$, $i \neq 2$ or $j \neq 3$.

The results using these constraints are depicted in Figure 8.

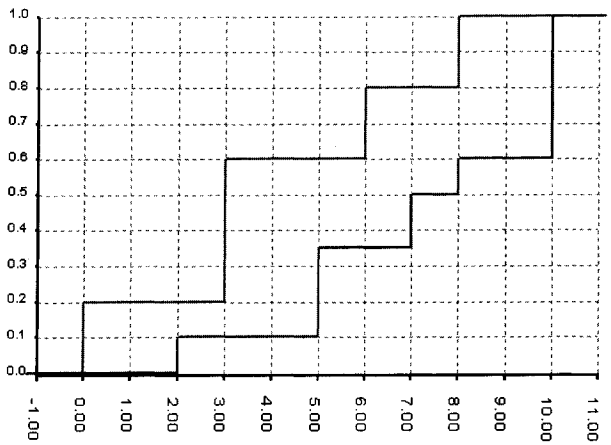



Figure 8. $\mathbf{F}_z(.)$, where the mode point is in $\mathbf{z}_{23}$.

Notice that the envelopes in Figure 8 are closer together than if the extra constraints are not present (as in Figure 3).

## Conditional unimodality constraint

Here we examine another related, but somewhat different situation: conditional unimodality. In this situation, the joint distribution is known to be unimodal for $x$ given a value for $y$, or unimodal for $y$ given a value for $x$.

For example, suppose that given some value $y$ of $Y$ in $\mathbf{y}_2$=[3, 4] in Table 3, the maximum density of the PDF $f_x(x|y)$ is at some value of $x \in \mathbf{x}_3$. Then, the average probability density in the cell with probability $p_{23}$ is greater than the average probability density in any cell with probability $p_{2k}$, $k \neq 3$. If the widths of intervals $\mathbf{z}_{2k}$ are the same, then $p_{23} \geq p_{2k}$, $k \neq 3$. In the more general case, the widths of the $\mathbf{z}_{2k}$ might not be the same. If width $w(\mathbf{z}_{23}) = c^* w(\mathbf{z}_{2k})$, then $p_{23} \geq c^* p_{2k}$. For the joint distribution tableau of Table 3, $w(\mathbf{z}_{21})=w(\mathbf{z}_{22})=w(\mathbf{z}_{23})$, so $p_{23} \geq p_{21}$

and $p_{23} \geq p_{22}$. These inequalities are constraints that, when included in the linear programming calls, will tend to squeeze the envelopes closer together than if these constraints were not included. Thus conditional unimodality can contribute constraints that tend to squeeze the envelopes bounding the CDF of $Z$.

Figure 9 shows the envelopes resulting from these new constraints. Notice that the envelopes are narrower than those of Figure 3, showing the narrowing influence of being able to assume conditional unimodality.

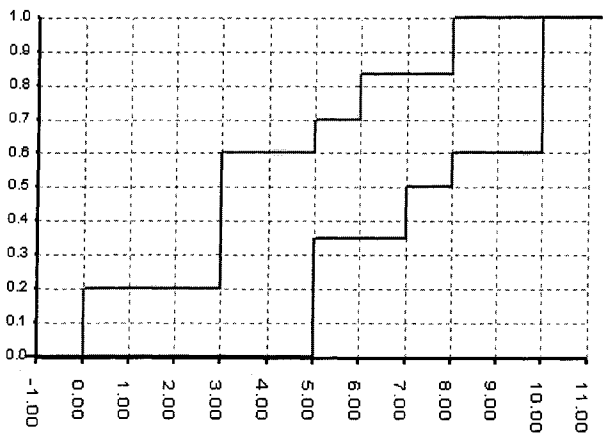



Figure 9. Conditional mode point in $\mathbf{z}_{23}$.

## Results and conclusion

In this paper, we present methods for using incomplete information about joint distributions to improve the envelopes around the CDF of a function of two marginals. More assumptions tend to give narrower result envelopes. More assumptions are good for improving results, but it is important that such assumptions are justified. We have shown that certain assumptions about the joint distribution of two marginals, that analysts will sometimes find useful and acceptable, can result in narrower CDF envelopes for functions of marginal random variables.

## References

[1] Berleant, D and J. Zhang. Representation and problem solving with the Distribution Envelope Determination (DEnv) method. *Reliability Engineering and System Safety* 85 (1-3) (2004a), pp. 153-168.

[2] Berleant, D and J. Zhang. Using correlation to improve envelopes around derived distribution. *Reliable Computing* 10 (2) (2004b), pp. 139-161.

[3] Berleant, D and J. Zhang. Bounding the times to failure of 2-component systems. *IEEE Transactions on Reliability*, 53 (4) (2004c), pp. 542-550.

[4] Clemen, R. and T. Reilly. Correlations and copulas for decision and risk analysis. *Management Science* 45 (2) (February 1999), pp. 208-224.

[5] Couso, I., S. Moral and P. Walley. Examples of independences for imprecise probabilities. *Proceedings of 1st International Symposium on Imprecise Probabilities and Their Applications*, Ghent, Belgium, 29 June – 2 July, 1999.

[6] Embrechts, P., F. Lindskog and A. McNeil. Modelling dependence with copulas and applications to risk management. In T. Rachev, *Handbook of Heavy Tailed Distributions in Finance*. Elsevier Science Ltd., 2003, pp. 329-384.

[7] Ferson, S. What Monte Carlo methods cannot do. *Journal of Human and Ecological Risk Assessment* 2 (4) (1996), pp. 990-1007.

[8] Ferson, S. And M. Burgman. Correlations, dependency bounds and extinction risks. *Biological Conservation* 73 (1995), pp. 101-105.

[9] Fetz, Th. and M. Oberguggenberger. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety* 85 (2004), pp. 73-87.

[10] Ingram, G.E., E.L. Welker, and C.R. Herrmann, "Designing for reliability based on probabilistic modeling using remote access computer systems," *Proc. 7th Reliability and Maintainability Conference*, American Society of Mechanical Engineers, 1968, pp. 492-500.

[11] Kolmogoroff (a.k.a. Kolmogorov), A., Confidence limits for an unknown distribution function, *Annals of Mathematical Statistics* 12 (4) (1941), pp. 461-463.

[12] Levy, H., ed., *Stochastic dominance: Investment Decision Making under Uncertainty*, Springer, New York, 1998.

[13] Nelsen, R. *An introduction to copulas*. Springer, New York, 1999.

[14] Neumaier, A. Clouds, fuzzy sets and probability intervals, *Reliable Computing* 10 (2004), 249-272.

[15] Williamson, R. And T. Downs. Probabilistic arithmetic i: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4 (1990).

[16] Zhang, J. And D. Berleant. Envelopes around cumulative distribution functions from interval parameters of standard continuous distributions. *Proceedings of North American Fuzzy Information Processing Society* (NAFIPS 2003), Chicago, pp. 407-412.

# CHAPTER 5. REPRESENTATION AND PROBLEM SOLVING WITH DISTRIBUTION ENVELOPE DETERMINATION (DENV)

A paper published in the Journal of Reliability Engineering and System Safety 85: 153-168, 2004.

Daniel Berleant and Jianzhong Zhang

## Abstract

Distribution Envelope Determination (DEnv) is a method for computing the CDFs of random variables whose samples are a function of samples of other random variable(s), termed inputs. DEnv computes envelopes around these CDFs when there is uncertainty about the precise form of the probability distribution describing any input. For example, inputs whose distribution functions have means and variances known only to within intervals can be handled. More generally, inputs can be handled if the set of all plausible cumulative distributions describing each input can be enclosed between left and right envelopes. Results will typically be in the form of envelopes when inputs are envelopes, when the dependency relationship of the inputs is unspecified, or both. For example in the case of specific input distribution functions with unspecified dependency relationships, each of the infinite number of possible dependency relationships would imply some specific output distribution, and the set of all such output distributions can be bounded with envelopes. The DEnv algorithm is a way to obtain the bounding envelopes. DEnv is implemented in a tool which is used to solve problems from a benchmark set.

**Keywords.** DEnv, p-boxes, aleatory uncertainty, epistemic uncertainty, second order uncertainty, uncertainty quantification, 2nd order uncertainty, reducible uncertainty, imprecise probabilities, challenge problems, envelopes, derived distributions, Statool.

## Introduction

The DEnv (Distribution Envelope Determination) algorithm is a method for computing distributions whose samples are some function of the samples of other input distributions, even under non-traditional conditions of severely limited knowledge about the inputs.

Under traditional conditions of known dependency relationships among precisely defined input distributions, solutions based around Monte Carlo simulation have an extensive literature, although MC gives results that are potentially problematic (Ferson 1996 [9]) and whose interpretation can be complicated by random variation especially in tails and other unlikely regions of system behavior. A large literature also addresses analytical solutions, which tend to require certain well-defined classes of distributions as inputs (Springer 1979 [28] is fairly comprehensive up to its time of writing). When the input random variables to be combined are independent in the traditional sense that the probability of a joint event is the product of the probabilities of its constituent events (often termed stochastic [7] or statistical independence), solutions based on numerical convolution are well known (Ingram et al. 1968 [16], Colombo and Jaarsma 1980 [6], Kaplan 1981 [17], Moore 1984 [20]). Lodwick's (2003 [19]) method is applied to multivariate examples with repeating variables and stated to be usable when variables are non-independent, or when their dependencies are unspecified, which is among the following problem characteristics that pose a challenge to traditional approaches.

(1) Sample values of one of the random variables may be described by a distribution, while sample values of another may be known only to within an interval.

(2) The input random variables may not be independent, and their dependency relationship may be unknown or only partly known. We will use the term *unknown*

*dependency* to describe this situation. The term "unknown interaction" has also been used (Couso et al. 1999).

(3) There may be insufficient information available to assign a specific distribution to an input random variable.

The problem of combining a distribution with an interval, (1) above, was addressed by Berleant (1993 [1]). When input dependencies are unknown, (2) above, the result random variable cannot in general be described with a single distribution, because each possible dependency relationship between the inputs leads to its own result distribution. Frank et al. 1987 [14] discuss the distribution of sums and products of samples of other distributions under this condition. Envelopes, also called p(robability)-bounds or p(robability)-boxes (Ferson et al. [10]) can be found which surround the family of possible result distributions. If these envelopes around the results are to be used in turn as inputs to produce further results, the algorithm for obtaining the further results must be able to use envelopes as inputs. This is also the problem of (3) above. We review solutions to (2), and then (3), next.

One approach to manipulating envelopes and distributions with unknown dependency relationships is based on the Probabilistic Arithmetic of Williamson and Downs (1990 [30]), which in turn is built on a foundation of copulas (Nelsen 1999 [21]). Probabilistic Arithmetic is one component of the commercially available RiskCalc (Ferson 2002 [8]) software. An approach based on sets of probability measures was applied to problems from a benchmark set (Oberkampf et al., this issue [23]) by Fetz and Oberguggenberger (this issue [13]), as was a Monte-Carlo based approach (Red-Horse this issue [24]). Ferson and Hajagos (this issue [11]) also address the problems using the just-mentioned Probabilistic Arithmetic. Tonon (this issue [29]) addresses Problem B using random set theory. Further solutions and insights were also presented by others at a recent workshop (Sandia 2002 [26]). In related work Neumaier [22] recently described clouds, a concept capable of expressing and manipulating families of CDFs bounded by left and right envelopes. The approach described in this paper

is Distribution Envelope Determination (DEnv), which relies on safely discretized distributions and linear programming. It has been reported on a theoretical basis (Berleant and Goodman-Strauss 1998 [2]) and implemented in a tool (Berleant et al. 2003 [3]). Applications have also been described (Berleant et al. 2002 [4]; Sheblé and Berleant 2002 [27]).

Equivalence properties of DEnv, Probabilistic Arithmetic, imprecise probabilities, and Dempster-Shafer structures are described by Regan et al. [25]. It appears these approaches are largely equivalent in their ability to construct envelopes around cumulative distributions in the real domain. They are also extendable to fuzzy numbers. Questions about how they compare in terms of computational speed and in ability to express and use inputs that are in non-cumulative form have still not been fully resolved. We feel that DEnv has an advantage in understandability compared to other methods. For example Probabilistic Arithmetic requires an understanding of copulas. Random sets also require specialized knowledge. Although DEnv uses linear programming (LP), knowledge of LP is widespread, and in DEnv may be viewed as a black box.

## Concise review of the DEnv algorithm

This section describes DEnv concisely and abstractly. Section 2 covers DEnv less formally, but at length and in detail in the context of a set of challenge problems [23]. The reader may choose to skip directly to section 2 without loss of continuity and refer back to this section later as needed, may choose to use this section as a foundation, or may take some intermediate course.

DEnv begins with two inputs, probability density functions $f_x(.)$ and $f_y(.)$ describing samples $x$ and $y$ of random variables $X$ and $Y$. DEnv will characterize the CDF (cumulative distribution function) $F_z(z)$ of samples $z=g(x,y)$ of random variable $Z$, given function $g$. The input PDFs $f_x(.)$ and $f_y(.)$ are discretized by partitioning the support (i.e. the domain over

which a PDF is non-zero) of each, yielding intervals $\mathbf{x}_i$, $i$=1…$I$, and $\mathbf{y}_j$, $j$=1…$J$. Each $\mathbf{x}_i$ is assigned a probability mass $p_{\mathbf{x}_i} = p(x \in \mathbf{x}_i) = \int_{x_0=\underline{\mathbf{x}_i}}^{\overline{\mathbf{x}_i}} f_x(x_0)dx_0$, where interval-valued symbols are shown in bold, and interval $\mathbf{x}_i$ has lower bound $\underline{\mathbf{x}_i}$ and upper bound $\overline{\mathbf{x}_i}$. Similarly, each $\mathbf{y}_j$ is assigned a probability mass $p_{\mathbf{y}_j} = p(y \in \mathbf{y}_j) = \int_{y_0=\underline{\mathbf{y}_j}}^{\overline{\mathbf{y}_j}} f_y(y_0)dy_0$. The $\mathbf{x}_i$'s and $\mathbf{y}_j$'s and their probabilities form the marginals of a discretized joint distribution called a *joint distribution tableau* (Table 1), the interior cells of which each contain two items. One is a probability mass $p_{ij} = p(x \in \mathbf{x}_i \wedge y \in \mathbf{y}_j)$. If the value of $x$ gives no information about the value of $y$, and vice versa, then $x$ and $y$ are independent and $p_{ij} = p(x \in \mathbf{x}_i) \cdot p(y \in \mathbf{y}_i) = p_{\mathbf{x}_i} \cdot p_{\mathbf{y}_j}$, where $p_{\mathbf{x}_i}$ is defined as $p(x \in \mathbf{x}_i)$ and $p_{\mathbf{y}_j}$ as $p(y \in \mathbf{y}_j)$. The second item is an interval that bounds the values $z$=$g(x,y)$ may have, given $x \in \mathbf{x}_i \wedge y \in \mathbf{y}_j$. In other words, $\mathbf{z}_{ij}$=$\mathbf{g}(\mathbf{x}_i,\mathbf{y}_j)$.

| $x\rightarrow$<br>$y\downarrow$<br>$z=g(x,y)\searrow$ | | $\mathbf{x}_i$<br>$p_{\mathbf{x}_i} = p(x \in \mathbf{x}_i)$ | |
|---|---|---|---|
| ⋮ | | ⋮ | |
| $\mathbf{y}_j$<br>$p_{\mathbf{y}_j} = p(y$ | | $\mathbf{z}_{ij}$=$\mathbf{g}(\mathbf{x}_i,\mathbf{y}_j)$<br>$p_{ij} = p(x \in \mathbf{x}_i \wedge y$ | |
| ⋮ | | ⋮ | |

Table 1. General form of a joint distribution tableau.

To better characterize the CDF $F_z(.)$, we next convert the set of interior cells of the joint distribution tableau into cumulative form. Because the distribution of each probability mass $p_{ij}$ over its interval $\mathbf{z}_{ij}$ is not defined by the tableau, values of $F_z(.)$ cannot be computed precisely. However they can be bounded, so DEnv does this by computing the interval-valued function $\mathbf{F}_z(.)$ as

$$\underline{\mathbf{F}_z}(z_0) = \sum_{i,j|\underline{\mathbf{z}_{ij}} \le z_0} p_{ij} \text{ and } \overline{\mathbf{F}_z}(z_0) = \sum_{i,j|\underline{\mathbf{z}_{ij}} \le z_0} p_{ij}, \quad (1)$$

resulting in right and left envelopes respectively for $\mathbf{F}_z(.)$.

An additional complication occurs if the dependency relationship between $x$ and $y$ is unknown, because then the values of the $p_{ij}$'s are underdetermined and so Equations (1) cannot be evaluated. However, the $p_{ij}$'s in a column of a joint distribution tableau must sum to $p_{\mathbf{x}_i}$ and the $p_{ij}$'s in a row must sum to $p_{\mathbf{y}_j}$ giving three sets of constraints: $p_{\mathbf{x}_i} = \sum_j p_{ij}$, $p_{\mathbf{y}_j} = \sum_i p_{ij}$, and $p_{ij} \ge 0$, for $i=1\ldots I$, $j=1\ldots J$. These constraints are all linear, and so may be passed to a linear programming routine. Linear programming takes as input linear constraints on variables, which in this case are the $p_{ij}$'s, and an expression in those variables to minimize, for example, $\underline{\mathbf{F}_z}(z_0) = \sum_{i,j|\underline{\mathbf{z}_{ij}} \le z_0} p_{ij}$ for some given value $z_0$. The output would then be the minimum value that $\underline{\mathbf{F}_z}(z_0)$ could have such that the values assigned to the $p_{ij}$'s are consistent with the constraints. $\overline{\mathbf{F}_z}(z_0) = \sum_{i,j|\underline{\mathbf{z}_{ij}} \le z_0} p_{ij}$ is maximized similarly. These envelopes are less restrictive (i.e. farther apart) than when the $p_{ij}$'s are determined by an assumption of independence or some other dependency relationship so that linear programming is not needed.

These ideas generalize naturally to $n$ marginals, which would require an $n$-dimensional joint distribution tableau (Section 2.6).

## The challenge problems and the DEnv technique

In this section the DEnv technique is explained in the context of the challenge problems given by Oberkampf et al. [23]. Solutions are presented and explained for all of the challenge problems, which include six scenarios involving computation of $y=(a+b)^a$ and also a spring problem.

## Problem 1: setting the stage

Problem 1 is to find the range of $y=(a+b)^a$ given $a\in[0.1, 1]$ and $b\in[0, 1]$. The minimum for $y$ occurs when $a=0.37$, which is not an endpoint of the interval for $a$, and when $b=0$, leading to the answer $y\in[0.69, 2]$. Using only endpoints of input intervals to compute bounds on result intervals can, as in this case, generate misleading results. This is a well-studied issue in interval computing and occurs numerous times in the challenge problems with respect to intervals for $a$ and in the spring problem.

Challenge Problems 2-6 have, as givens, one or more sources of information about $a$ and also about $b$. The sources often are specified to have equal credibility. The equal credibility stipulation contains significant ambiguity. This has serious implications for the solutions, which are discussed later in Section 3. In the present section we seek solutions for the problems and hence must precisely define them. Therefore we resolve the ambiguity by modeling credibility using probability.

When different information sources have equal credibility we interpret this to mean that the actual but unknown value has the same probability of being binned in the interval of one information source as it does of being binned in the interval of any other. This interpretation allows different information sources to have equal credibility while providing probabilities for intervals that are disjoint, nested, or overlapping (all of which could occur in real situations). Which of those situations occurs in a given problem has little effect on the solution using DEnv, an algorithm which is consistent with standard properties of probability and requires, to be applied, that a problem be modeled using intervals and associated probabilities. (Another typical way of modeling problems for processing by DEnv is to discretize probability distributions, as occurs in the solution to Challenge Problem 6.)

## Problem 2: $a \in [0.1, 1]$ with equally credible intervals for $b$

The following facts about this problem are defined by Oberkampf et al. (this issue [23]) and the discussion above: $y=(a+b)^a$, with $a \in [0.1, 1]$ and $p(b \in \mathbf{b}_1)=p(b \in \mathbf{b}_2)=p(b \in \mathbf{b}_3)=p(b \in \mathbf{b}_4)$ for intervals $\mathbf{b}_1 \ldots \mathbf{b}_4$, each provided by a different information source all of which are equally credible.

Table 2 shows, for Problem 2a, the givens for $a$ and $b$ and the results that follow using interval arithmetic to get intervals describing the consequent range of $y$. For the last row, the intervals for $a$, $b$, and hence $y$ are the same as for Problem 1 above and hence require the same attention to interior points of $a$.

| Intervals given for $b$ | Intervals for $y=(a+b)^a$, given $a \in [0.1, 1]$ |
|---|---|
| $\mathbf{b}_1=[0.6, 0.8]$<br>$p=0.25$ | $\mathbf{y}_1=[0.96, 1.8]$<br>$p_1=0.25$ |
| $\mathbf{b}_2=[0.4, 0.85]$<br>$p=0.25$ | $\mathbf{y}_2=[0.9, 1.85]$<br>$p_2=0.25$ |
| $\mathbf{b}_3=[0.2, 0.9]$<br>$p=0.25$ | $\mathbf{y}_3=[0.81, 1.9]$<br>$p_3=0.25$ |
| $\mathbf{b}_4=[0, 1]$<br>$p=0.25$ | $\mathbf{y}_4=[0.69, 2]$<br>$p_4=0.25$ |

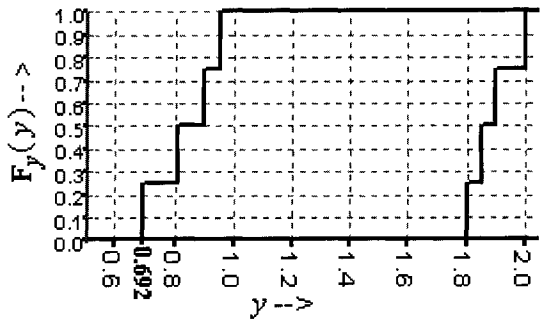
Table 2. The interval given by Problem 2a about the value of $a$ (top right cell), intervals given about the value of $b$ (left column), and the implications of those intervals for the value of $y$ (right column).

Figure 1 (top) shows $y$ in Problem 2a graphically. The distribution envelopes shown in the graphs of Figure 1 may be derived straightforwardly from the right-hand column of their corresponding tables as follows.

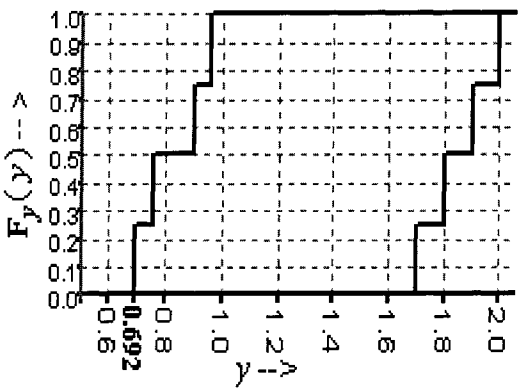
1) *Left envelope.* This envelope, $\overline{\mathbf{F}}(y)$, represents the maximum possible cumulation of probability mass for any given value of $y$. It is obtained under the extreme assumption that the probability mass associated with each interval is distributed as an impulse at the low bound of the interval. (Such an extreme assumption is permitted because an interval does not imply anything about how its probability mass is distributed beyond that it is distributed within its bounds.) Therefore any interval whose low bound is at or below a value $y_0$ can contribute up to its full probability mass to the cumulation of $y$ at $y_0$, while other intervals cannot contribute any mass. Thus for each of Problems 2a-2c, $\overline{\mathbf{F}}(y) = \sum_{k|\underline{\mathbf{y}}_k \le y} p_k$, where bold signifies interval-valued symbols, underlining refers to an interval's low bound, and overlining refers to an interval's high bound.

2) *Right envelope.* This envelope, $\underline{\mathbf{F}}(y)$, represents the minimum cumulation of probability mass for any given value of $y$. This is obtained under the extreme assumption that each mass is distributed as an impulse at the high bound of its corresponding interval. Thus, $\underline{\mathbf{F}}(y) = \sum_{k|\overline{\mathbf{y}}_k \le y} p_k$.

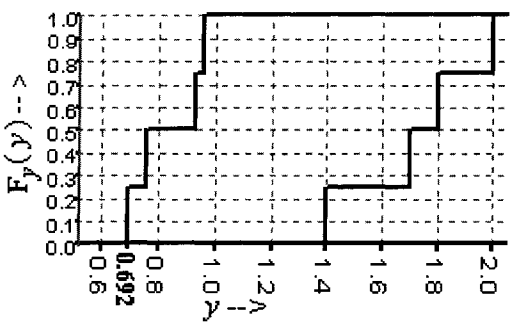
Figure 1 also shows graphs for $y$ in Problems 2b and 2c, derived as just described. (For Problem 2a, the intervals $\mathbf{y}_k$ were shown in Table 2, while envelopes around the CDF for $b$ appear later in Figure 15; for Problem 2c, intervals $\mathbf{y}_k$ appear later in Table 4 and envelopes around the CDF of $b$ appear later in Figure 13, top.) That the intervals for $b$ in Problem 2a are nested, in 2b are overlapping, and in 2c are completely disjoint [23] does not affect the process of computing intervals and graphs for $y$.




**(Problem 2a↑)**




**(Problem 2b↑)**




**(Problem 2c↑)**

Figure 1. Envelopes around the cumulative distribution of the value of $y$ in Problems 2a, 2b, and 2c implied by the intervals $y_k$ for each problem.

## Problem 3: intervals for *a* and intervals for *b*

In the preceding problem four sources of information about $b$ led to computing, then combining, four cases for $y$. In this problem, four cases of $b$ for each of the three cases for $a$ lead to 12 cases for $y$. Given equal probability assignments for each case of $a$, and likewise for $b$, if $a$ and $b$ are assumed independent in the sense used throughout this paper, that a sample of one gives no information about the other, then the 12 cases for $y$ will each have equal probability. (Fetz and Oberguggenberger [13] show the implications of different kinds of independence [7] for the challenge problems.) If $a$ and $b$ are not independent then cases for $y$ will typically have different probabilities.

The format of Table 2 can be generalized to express these situations. Figure 2 shows both the table and resulting envelopes for Problem 3a, and the envelopes for 3b (3c will be discussed in detail later). It is convenient to describe the tables using the following terminology.

- The leftmost column and topmost row describe $a$ and $b$, and are called *marginals.*
- Cells of the table with intervals labeled $\mathbf{y}_{ij}$ are called *interior cells.*
- The distribution of the probability mass of a cell over its interval is called the cell's *mini-distribution.*
- The entire table is called a *joint distribution tableau.*

The intervals and probability masses in the marginals discretize $a$ and $b$. The marginal intervals and the interior cell probability masses together discretize the joint distribution of $a$ and $b$. The interior cell probability masses and their intervals give a discretization of the distribution of $y=(a+b)^a$. This was shown abstractly in Table 1.

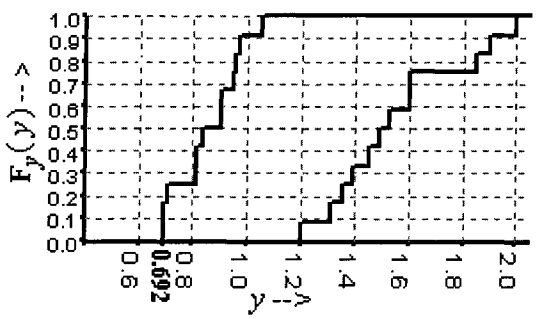
Note that different cells in the same marginal can have overlapping intervals, as in Table 2 and Figure 2 (top). A simple example illustrates the meaning of overlaps. Consider a symmetric probability density function (PDF) with support over [0, 4]. A very coarse

discretization consists of the single interval [0, 4] with an associated probability mass of 1, because the PDF contains a mass of 1 distributed over the interval [0, 4]. Since the PDF is symmetric, it may also be discretized as two probability masses, one of 0.5 distributed appropriately over [0, 2], and another also of 0.5 and also distributed appropriately over (2, 4] (assuming the PDF does not have an impulse at exactly 2). Yet, the wider and overlapping intervals [0, 3] and [1, 4] can also give a valid discretization of the same PDF. One way to do that is to specify the same probability mass and mini-distribution for [0,3] that was just used for [0, 2] (implying that no mass assigned to [0, 3] happens to be distributed above 2), and the same mini-distribution over [1, 4] that was just used for (2, 4] (implying that no mass assigned to [1, 4] is distributed at or below 2), resulting in exactly the same mini-distributions as in the case of the {[0, 2], (2, 4]} discretization. As a final example consider a discretization with extreme overlap consisting of two intervals, each with range [0, 4] and probability mass 0.5. It is certainly possible to distribute one mass within [0, 4], then add in the other mass, distributed appropriately within the same interval, to get the original PDF.

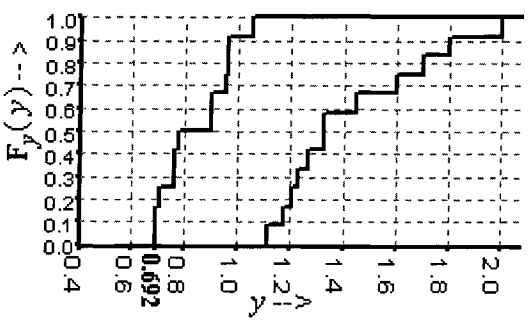
Just as a marginal of a joint distribution tableau discretizes the distribution of an input random variable, the interior cells of the tableau collectively discretize the distribution of $y=(a+b)^a$ even though the phenomenon of overlapping result intervals is often present (e.g. Table 2 and Figure 2, top) whether or not there are overlapping intervals in either marginal.

| $a\rightarrow$ | [0.5, 0.7] | [0.3, 0.8] | [0.1, 1] |
|---|---|---|---|
| $b\downarrow$ $y\searrow$ | $p$=0.33 | $p$=0.33 | $p$=0.33 |
| [0.6, 0.6] | $\mathbf{y}_{11}$=[1.0, 1.2] | $\mathbf{y}_{21}$=[0.97, 1.3] | $\mathbf{y}_{31}$=[0.96, 1.6] |
| $p$=0.25 | $p_{11}$=0.083 | $p_{21}$=0.083 | $p_{31}$=0.083 |
| [0.4, 0.85] | $\mathbf{y}_{12}$=[0.95, 1.4] | $\mathbf{y}_{22}$=[0.90, 1.5] | $\mathbf{y}_{32}$=[0.9, 1.85] |
| $p$=0.25 | $p_{12}$=0.083 | $p_{22}$=0.083 | $p_{32}$=0.083 |
| [0.2, 0.9] | $\mathbf{y}_{13}$=[0.84, 1.4] | $\mathbf{y}_{23}$=[0.81, 1.5] | $\mathbf{y}_{33}$=[0.81, 1.9] |
| $p$=0.25 | $p_{13}$=0.083 | $p_{23}$=0.083 | $p_{33}$=0.083 |
| [0, 1] | $\mathbf{y}_{14}$= [0.71, 1.4] | $\mathbf{y}_{24}$= [0.69, 1.6] | $\mathbf{y}_{34}$= [0.69, 2] |
| $p$=0.25 | $p$=0.083 | $p$=0.083 | $p$=0.083 |

Joint distribution tableau for Problem 3a (numbers are to 2 significant digits).




Envelopes enclosing the CDF for $y$ in Problem 3a.




Envelopes enclosing the CDF for $y$ in Problem 3b.

Figure 2. The joint distribution tableau for Problem 3a (top) was used to generate envelopes for $y$ (middle) using $\overline{\mathbf{F}}(y_0) = \sum_{i,j|\mathbf{y}_{ij} \le y_0} p_{ij}$ for the left envelope and $\underline{\mathbf{F}}(y_0) = \sum_{i,j|\overline{\mathbf{y}}_{ij} \le y_0} p_{ij}$ for the right envelope (see Section 1.1). Envelopes for $y$ in Problem 3b are also shown (bottom).

## Removing the independence assumption

Figure 2 assumes the marginals are independent in the standard statistical sense that the probability assigned to each interior cell in a joint distribution tableau is the product of its marginal cell probabilities. For example, consider the cell in the lower right corner of the tableau of Figure 2 (top) and the marginal cells for its row and its column. The probability that $a$ is in [0.1, 1] and is binned in the marginal cell with that interval rather than in any other marginal cell whose interval it might be consistent with, is specified as 0.33. Similarly the probability that $b$ is in the marginal cell with interval [0, 1] is 0.25. Therefore the probability assigned to the lower right interior cell is $0.25 \cdot 0.33 = 0.083$.

All interior cell probabilities were computed similarly. However if the marginals are not independent then the interior cell probabilities are determined by the details of the dependency relationship, whatever it is. A human analyst could for example manually type in interior cell probabilities to express some particular dependency relationship, or software could fill them in based on some formula defining a dependency relationship. Equations (1) and the equations in the caption of Figure 2 still apply. The DEnv technique can also be extended to the case where the dependency between $a$ and $b$ is not determined. This case has the following two subcases: (*i*) the dependency between $a$ and $b$ is unknown, discussed next, and (*ii*) partial knowledge about their dependency exists, discussed after.

**Unknown dependency.** If the dependency relationship is not specified then the interior cell probabilities are not determined. They are however constrained by the marginal cell probabilities. The probability in each marginal cell in the left column is distributed

among the interior cells in its row. Also the probability in each marginal cell in the top row is distributed among the interior cells in its column. Thus in Table 3 there are four row constraints and three column constraints.

| $a\rightarrow$ | $\mathbf{a}_1$=[0.8,1] | $\mathbf{a}_2$=[0.5,0.7] | $\mathbf{a}_3$=[0.1,0.4] |
|---|---|---|---|
| $b\downarrow$ $y\searrow$ | $p$=0.33 | $p$=0.33 | $p$=0.33 |
| $\mathbf{b}_1$=[0.8,1] | [1.5,2] | [1.1,1.45] | [0.99,1.1] |
| $p$=0.25 | $p_{11}$=? | $p_{21}$=? | $p_{31}$=? |
| $\mathbf{b}_2$=[0.5,0.7] | [1.2,1.7] | [1.0,1.3] | [0.93,1.0] |
| $p$=0.25 | $p_{12}$=? | $p_{22}$=? | $p_{32}$=? |
| $\mathbf{b}_3$=[0.1,0.4] | [0.92,1.4] | [0.78,1.1] | [0.76,0.93] |
| $p$=0.25 | $p_{13}$=? | $p_{23}$=? | $p_{33}$=? |
| $\mathbf{b}_4$=[0,0.2] | [0.84,1.2] | [0.71,0.93] | [0.69,0.89] |
| $p$=0.25 | $p_{14}$=? | $p_{24}$=? | $p_{34}$=? |

| **Row constraints** | **Column constraints** |
|---|---|
| $0.25=p_{11}+p_{21}+p_{31}$ | $0.33=p_{11}+p_{12}+p_{13}+p_{14}$ |
| $0.25=p_{12}+p_{22}+p_{32}$ | $0.33=p_{21}+p_{22}+p_{23}+p_{24}$ |
| $0.25=p_{13}+p_{23}+p_{33}$ | $0.33=p_{31}+p_{32}+p_{33}+p_{34}$ |
| $0.25=p_{14}+p_{24}+p_{34}$ | |

Table 3. Joint distribution tableau for Problem 3c (top), expressing the case of unknown dependency between $a$ and $b$. Thus the interior cell probabilities are underdetermined. They are however partially constrained by the marginal cells, each of which defines a row or column constraint (bottom).

The restrictions on the interior cell probabilities imposed by the row and column constraints are the key to finding the height of the left or right envelope at a given value of $y=(a+b)^a$. The formula for the left envelope is

$$\overline{\mathbf{F}_y}(y_0) = \sup_{p_{ij}, i=1\ldots I, j=1\ldots J | C} \sum_{i,j | \underline{\mathbf{y}_{ij}} \le y_0} p_{ij} \quad \text{and for the right envelope,}$$

$$\underline{\mathbf{F}_y}(y_0) = \inf_{p_{ij}, i=1\ldots I, j=1\ldots J | C} \sum_{i,j | \mathbf{y}_{ij} \le y_0} p_{ij} \qquad (2)$$

where $C$ refers to the set of row and column constraints that must hold. In other words, the sup operation finds probability mass assignments for the $p_{ij}$'s that give the maximum value for the summation that is possible while maintaining consistency with constraint set $C$, and the inf operation acts similarly to give the minimum value. These equations are like Equations (1) augmented with sup and inf, and express the variability of the $p_{ij}$'s as constrained by the row and column constraints. The intuitions behind these formulas were reviewed toward the end of Section 1.1, and are detailed along with some other salient points next.

*Left envelope.* The height of the left envelope at some value $y=y_0$ on the horizontal axis is the maximum possible cumulation of probability mass over the interval $(-\infty, y_0)$. This may be obtained as follows.

(1) Identify all interior cells with interval low bounds at or below $y_0$. Each potentially contributes its entire probability mass to the cumulation at $y_0$, because its mini-distribution can be specified so as to distribute its entire mass over values at or below $y_0$. Other interior cells cannot distribute any of their probability masses to values at or below $y_0$ no matter what their mini-distributions are, because their intervals' low bounds are above $y_0$.

*Example* 1. In Table 3, for $y$=0.95 the maximum cumulation will involve:

$p_{32}, p_{13}, p_{23}, p_{33}, p_{14}, p_{24}, \& p_{34}$.

(2) Maximize the sum of the probability masses of the previously identified interior cells. This may be done by using linear programming as a software subroutine call and passing in as inputs, (*i*) the constraints defined by the row and column constraints, and (*ii*) the function to maximize. For *Example* 1 above, the function to maximize is therefore $p_{32}+p_{13}+p_{23}+p_{33}+p_{14}+p_{24}+p_{34}$. Linear programming finds values for the $p_{ij}$ which maximize that function while satisfying the constraints. Maximization can actually be done manually by careful inspection, pushing masses around in the tableau, but using a computer to solve this as a linear programming program is more practical.

*Right envelope.* To find the height of the right envelope at $y=y_0$, instead of maximizing a sum of interior cell probability masses we minimize, because the right envelope expresses the minimum possible cumulation at each value of $y$. The interior cells whose pooled probability mass is to be minimized are those whose intervals have high bounds at or below $y_0$, because the full mass of each of those cells must be in the cumulation at $y_0$, although from the perspective of minimization we would wish otherwise. It is possible for the mini-distributions of other interior cells to be specified so as to allocate all of their respective masses above $y_0$, thereby not contributing to the cumulation. (An alternative to minimizing the mass of this set of interior cells is to maximize the mass of its complement.)

Having explained how to get the height of the left and right envelopes for a given value of $y$ we must now choose the values of $y$ at which to do this computation. For the left (right) envelope these values are the low (high) bounds of the interior cell intervals, because it is at these bounds that the envelope heights can change, since the maximization

(minimization) process depends on these bounds as described above and in Equations (1). Figure 3 shows the envelopes for Problems 3a and 3b when $a$ and $b$ have unknown dependency.

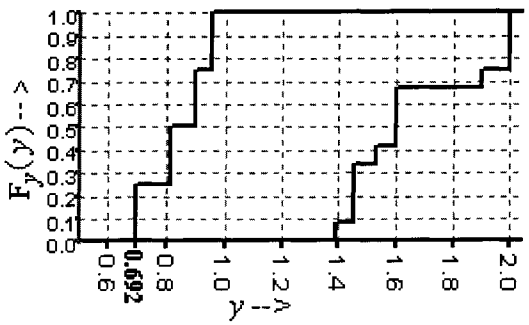



Figure 3. The envelopes shown here are more widely separated than those of Figure 2, because removing the independence assumption tends to weaken the results.

**Correlation.** Independence is a strong assumption. Simply removing that assumption leaves no information at all about the dependency relationship between $a$ and $b$, significantly weakening results as shown in Figure 3. An intermediate case exists when $a$ and $b$ are assumed correlated. Intuitively if $b$ is likely to be low when $a$ is low, and high when $a$ is high, then $a$ and $b$ are positively correlated. Alternatively if $b$ is likely to be high when $a$ is low, and low when $a$ is high, then $a$ and $b$ are negatively correlated. Consequently in a

tableau like that of Table 3 (top), if probability mass is concentrated in interior cells along a diagonal northwest-southeast path the correlation will be high, while if mass is concentrated along the other diagonal correlation will be low. More generally, consider a standard equation for the (Pearson) correlation coefficient $\rho$ in terms of expectations $E(.)$.

$$\rho = \frac{E(ab) - E(a)E(b)}{\sqrt{\left(E(a^2) - E(a)^2\right)\left(E(b^2) - E(b)^2\right)}}$$

The only term in this formula that is influenced by the joint distribution of $a$ and $b$ is $E(ab)$. The other terms depend only on the marginals, $a$ and $b$, whose descriptions are givens in the problems examined in this paper. Observe from the equation that higher values of $E(ab)$ imply higher values of $\rho$ relative to lower values of $E(ab)$. This is illustrated by the fact that if $a$ has two possible values $a_0$=9 and $a_1$=10, each with a 50% chance of occurring, and $b$ has the same distribution for its possible values $b_0$ and $b_1$, then if $a_0$ always co-occurs with $b_0$, and $a_1$ with $b_1$, satisfying the intuitive concept of high correlation, $E(ab)$=(9*9+10*10)/2=90.5. This is higher than if $a_0$ always co-occurs with $b_1$, and $a_1$ with $b_0$, in which case $a$ and $b$ satisfy the intuitive concept of low correlation and $E(ab)$=90. The difference is often more pronounced, as for $a_0$= −1, $a_1$=1, $b_0$= −1, and $b_1$=1.

If the term $E(ab)$ is assigned a value, range, minimum, or maximum, this can be used as a constraint to augment the row and column constraints. This will tend to restrict allocation of probability masses among the interior cells of a joint distribution tableau more than the row and column constraints alone (Berleant and Zhang, forthcoming [5]). For example, suppose in Challenge Problem 3c we state that $0.465 \leq E(ab)$. Then the row and column constraints in Table 3 would be augmented with a new constraint, derived next.

$0.46 \le \overline{\sum_{i,j} \mathbf{a}_i \cdot \mathbf{b}_j \cdot p_{ij}}$ (this is the constraint because we evaluate $E$(ab) from the interior cells of a joint distribution tableau, which produces an interval that is consistent with the requirement that $0.46 \le E(ab)$ if any part of the interval is 0.465 or more)

$= \sum_{i,j} \overline{\mathbf{a}_i \cdot \mathbf{b}_j \cdot p_{ij}}$ (the high bound of the sum is the sum of the high bounds)

$= \sum_{i,j} \overline{\mathbf{a}_i \cdot \mathbf{b}_j} \cdot p_{ij}$ (a number equals its high bound)

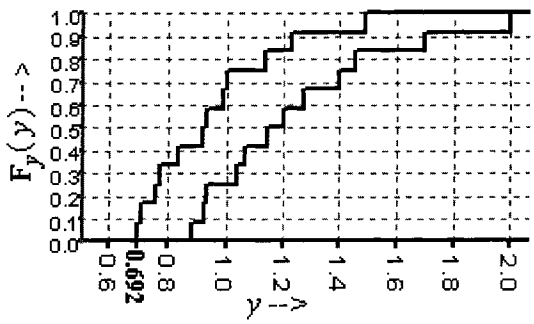
$= \sum_{i,j} \overline{\mathbf{a}_i} \cdot \overline{\mathbf{b}_j} \cdot p_{ij}$ (the high bound of the product of non-negative intervals is the product of the high bounds)

$$= 1 \cdot 1 \cdot p_{11} + 1 \cdot 0.7 \cdot p_{21} + 1 \cdot 0.4 \cdot p_{31} + 0.7 \cdot 1 \cdot p_{12} + 0.7 \cdot 0.7 \cdot p_{22} + 0.7 \cdot 0.4 \cdot p_{32} + 0.4 \cdot 1 \cdot p_{13} + 0.4 \cdot 0.7 \cdot p_{23} + 0.4 \cdot 0.4 \cdot p_{33} + 0.2 \cdot 1 \cdot p_{14} + 0.2 \cdot 0.7 \cdot p_{24} + 0.2 \cdot 0.4 \cdot p_{34} \ge 0.465$$

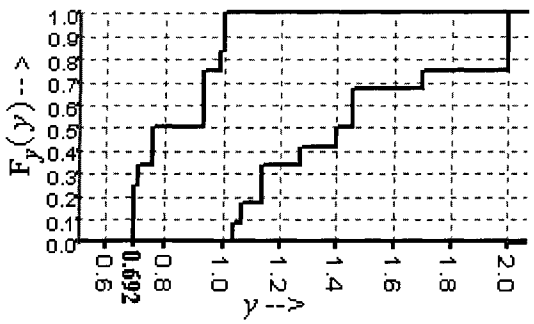
(which is read directly from the joint distribution tableau in Table 3, top, and is the new constraint to add to Table 3, bottom).

Note the assumptions $a \ge 0$, $b \ge 0$. To remove them, see [5].

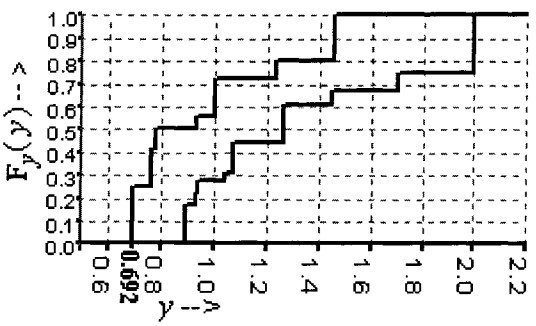
For the given marginal discretizations, the maximum value that can be attained for $E(ab)$ for any assignment of probability masses among the interior cells of the joint distribution tableau for Problem 3c without violating the row and column constraints is 0.47. Thus the constraint $E(ab) \ge 0.465$ enforces high correlation, which moves the envelopes closer together, excluding regions that the CDF can enter only when there is a significant chance of $a$ being when $b$ is low, or vice versa. This results in Figure 4 (iii). At the other extreme, the minimum value possible for $E(ab)$ for this problem is 0.081, so the constraint $E(ab) \le 0.09$ enforces low correlation, excluding regions that the CDF can enter only when there is a significant chance that $a$ and $b$ are both low or both high. This results in Figure 4 (iv). Note the significant differences in, for example, the shape of the right tail across the conditions of independence, low and high correlation.



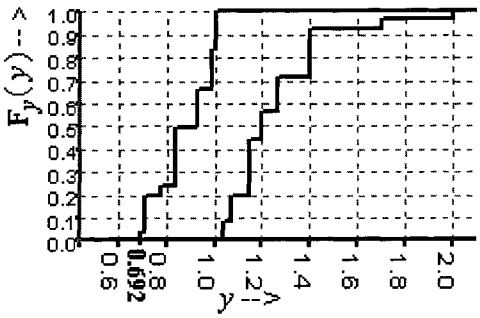

**(i) Independent inputs *a* and *b*.**




**(ii) Unknown dependency between *a* and *b*.**




**(iii) High correlation between *a* and *b*.**




**(iv) Low correlation between *a* and *b*.**

Figure 4. Solutions to Problem 3c in four variations (an additional variation will appear in Figure 14). Note that the unknown dependency condition (ii) results in envelopes that enclose those resulting from constrained dependency relationships such as those in the other three graphs.

## Problem 4: *a* is an interval and *b* is a left and right envelope pair

In this problem, $a$ is the same interval as in Problem 2. However, unlike in Problem 2, here $b$ is a family of lognormal distributions consisting of those distributions with means in [0, 1] and standard deviations in [0.1, 0.5]. The approach taken is to convert the family into a set of intervals with associated probability masses, thereby transforming Problem 4 into one like Problem 2, and then solve as in Problem 2. Therefore this section shows:

(1) how to convert a family of distributions into a set of intervals with associated probability masses,

(2) how to convert a listing of intervals for $(a+b)^a$ as in Section 2.1 into an equivalent joint distribution tableau, which is the standard format for representing problems to be solved with DEnv, and

(3) the solution for Challenge Problem 4.

### How to convert a family of distributions into a set of intervals and probability masses.

To apply DEnv, a set of intervals and their associated probability masses must be used to represent each marginal. To convert a family of PDFs into this form, the family is first represented as a pair of left and right envelopes that enclose the CDFs of the PDFs. Then these envelopes are converted into a set of intervals and a mass for each. A method is needed that does this and has the property that the resulting intervals and associated masses will, if converted back into envelopes, produce envelopes like the ones from which the intervals and masses were derived. One such method begins by tiling the space between the envelopes with rectangles such that the left side of each rectangle is on the left envelope, and the right side of each rectangle is on the right envelope. If the envelopes have a staircase shape this may be done exactly, while in the general case this will entail discretization since the rectangles have vertical sides. The span of a rectangle over the horizontal axis defines an interval, and the bottom-to-top height of the rectangle defines the probability mass for that interval. This gives a set of intervals $\mathbf{z}_k$, $k$=1…$K$, and their associated masses $p_k$. When these are converted to envelopes according to $\underline{\mathbf{F}}_z(z_0) = \sum_{k|\underline{\mathbf{z}}_k \le z_0} p_k$ and $\overline{\mathbf{F}}_z(z_0) = \sum_{k|\underline{\mathbf{z}}_k \le z_0} p_k$, (3) which are no more than re-subscripted forms of Equations (1), the result is envelopes identical to those from which the $\mathbf{z}_k$'s were derived, or similar if the original envelopes were not staircase shaped. Figure 5 shows an example. Some issues surrounding this are discussed further in Section 3.

## How to convert a 2-column table of intervals and probability masses into an equivalent joint distribution tableau.

A 2-column listing like Table 2 gives the single interval for $a$, lists all the intervals for $b$, and for each gives the result interval for $y=(a+b)^a$. The equivalent joint distribution tableau follows directly. Table 4 gives an example. Note how in the joint distribution tableau form, the interior cell masses $p_{11}...p_{41}$ add up to 1.0 (the mass of its single marginal cell for $a$ in the top right corner), and each mass in $p_{11}...p_{41}$ equals ("adds up to") the mass of the marginal cell to its left, in accordance with the four row constraints and one column constraint implied by the tableau.

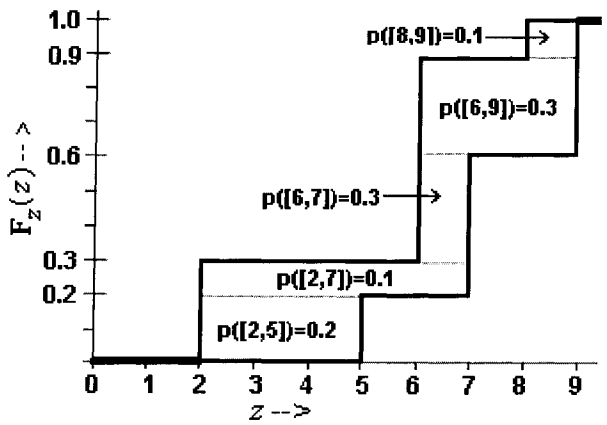



Figure 5. The left and right envelopes shown convert to the following intervals and probability masses: $p([2, 5])=0.2$, $p([2, 7])=0.1$, $p([6, 7])=0.3$, $p([6, 9])=0.3$, $p([8, 9])=0.1$. These intervals and masses, when converted back to envelopes using Equations (3), give envelopes identical to those shown.

| $b$ | $y=(a+b)^a$, given $a\in[0.1,1]$ |
|---|---|
| [0.6,0.8] $p$=0.25 | [0.96,1.8] $p$=0.25 |
| [0.5,0.7] $p$=0.25 | [0.93,1.7] $p$=0.25 |
| [0.1,0.4] $p$=0.25 | [0.76,1.4] $p$=0.25 |
| [0,1] $p$=0.25 | [0.69,2] $p$=0.25 |

**A 1-column table like Table 2 but for Problem 2c.**

| $a\rightarrow$ <br> $b\downarrow$ $y\searrow$ | [0.1,1] <br> $p$=1.0 |
|---|---|
| $b_1$=[0.6,0.8] $p$=0.25 | $y_{11}$=[0.96,1.8] $p_{11}$=0.25 |
| $b_2$= [0.5,0.7] $p$=0.25 | $y_{21}$= [0.93,1.7] $p_{31}$=0.25 |
| $b_3$= [0.1,0.4] $p$=0.25 | $y_{31}$= [0.76,1.4] $p_{31}$=0.25 |
| $b_4$= [0,1] $p$=0.25 | $y_{41}$= [0.69,2] $p_{41}$=0.25 |

**An equivalent joint distribution tableau.**

Table 4. Example (using Problem 2c for illustration instead of Problem 4, which would require many more rows) of conversion from the simple tabular format used in Section 2.1 to the joint distribution tableau format. In fact they are almost the same.

## The solution for challenge problem 4.

Monte Carlo simulation was used to generate envelopes enclosing the family of lognormal CDFs possible for $b$ that were specified by Problem 4. To do this, the specified intervals for mean $\mu$ and standard deviation $\sigma$ were each sampled, resulting in a fully specified lognormal CDF, the height of which was then evaluated at each of a predefined set of values of $b$. The evaluation process was repeated for additional pairs of samples of $\mu$ and $\sigma$ using the same values of $b$ and resulting in a set of CDF heights at each value of $b$. Then, for each value of $b$ the highest of the CDF heights was used as a point on the left envelope and the lowest was used as a point on the right envelope. The resulting envelopes were converted to a set of 59 intervals for $b$ as described earlier in this section. A joint distribution tableau was then constructed with 59 rows for $b$, and one column for the one interval provided for $a$ by the problem statement. Then, envelopes for the cumulative distribution of $y=(a+b)^a$ were constructed using DEnv just as in Problem 2. The results are shown in Figure 6.

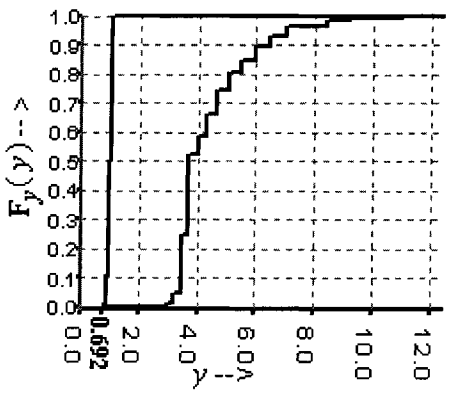



Figure 6. Solution to Problem 4.

## Problem 5: a set of intervals for *a* and a set of left and right envelope pairs for *b*

The new challenge posed by this problem is dealing with not one pair of envelopes describing $b$, but $n$ pairs. To handle this, we first combine the $n$ pairs of envelopes into a single composite pair. This pair can then be converted to intervals with associated probability masses as described in Section 2.3. At that point there will be a set of intervals and their masses for $b$, and the set of equally weighted intervals given by the problem definition for $a$. The problem can then be solved like Problem 3.

**Combining envelopes.** Problems 5a-5c are similar, in that each states three left-and-right pairs of envelopes of equal credibility. Therefore we assume a total cumulation of 0.333 for $b$ over $-\infty$ to $\infty$ for each of the three envelope pairs. To do this, each of the three pairs was normalized to reach a height of 0.333 instead of its original height of 1. Next, each pair of envelopes was converted into intervals and associated probability masses as in Section 2.3. At this point the sum of the masses of the intervals for each pair of envelopes is 0.333. Finally the entire set $J$ of all of the intervals from the three pairs of envelopes, with a combined mass of 1.0, were converted to a new combined pair of envelopes in accordance with Equations (3).

Having obtained a single, combined envelope pair, an equivalent set of intervals and associated probability masses can be derived as described in Section 2.3 (or set $J$ may simply be used), forming the marginal for $b$. The given intervals for $a$ are used to form the marginal for $a$. At this point the problem can be solved like Problem 3 in Section 2.2.

To solve Problem 5a we first used Monte Carlo simulation as in Section 2.3 to obtain a pair of envelopes for each source of information about $b$. These were combined as just described. The combined pair (see Figure 7) was converted into a set of intervals and associated probability masses as in Section 2.3, which was used as a marginal for a joint distribution tableau from which the result envelopes were derived. Problems 5b and 5c were

also solved this way. Figures 8-10 shows the results when $a$ and $b$ are independent, as well as when dependency is unknown. Note the scalloped envelopes in Figure 10, caused by gaps between the intervals given for $a$ and for standard deviation of $b$ [23].

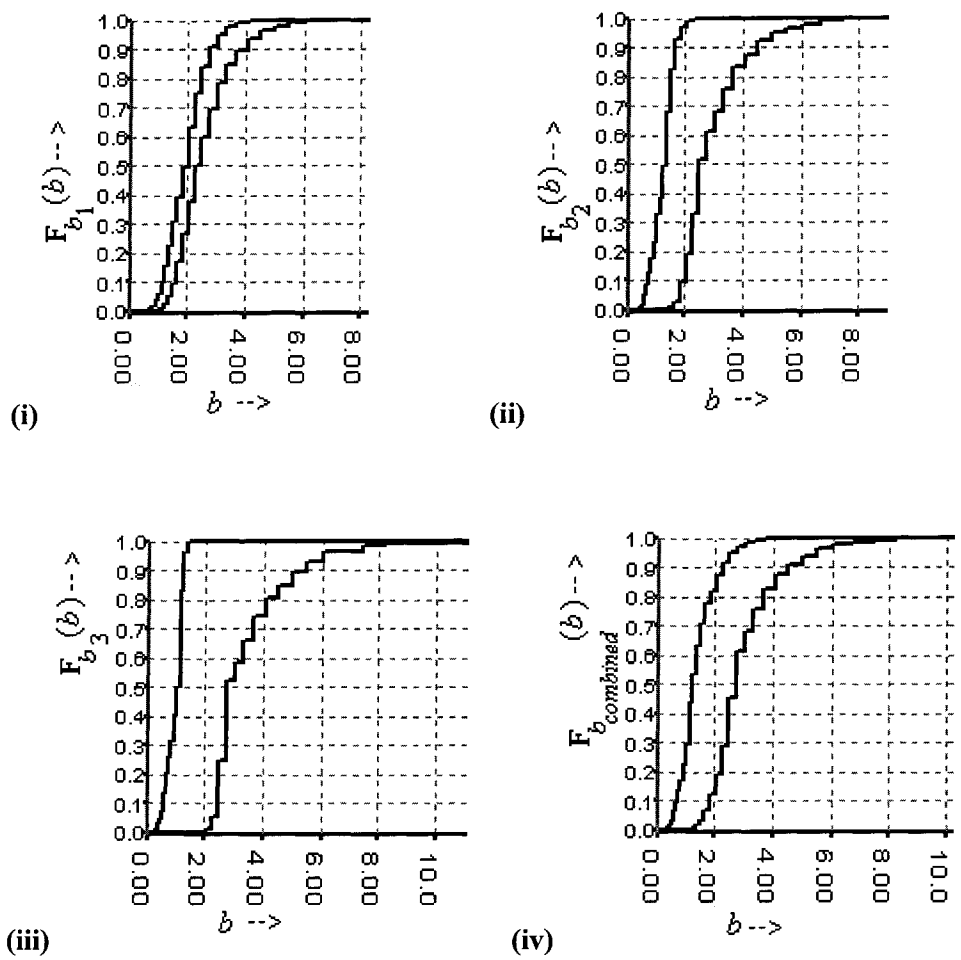



Figure 7. Three sources of information (i)-(iii) about $b$ given for Problem 5a (see [23]) and their combination (iv). Information $\mathbf{F}_{b_1}$ contains PDFs with (real-valued) means in [0.6, 0.8] and (real-valued) standard deviations in [0.3, 0.4]; (ii) $\mathbf{F}_{b_2}$ contains PDFs with means in [0.2, 0.9] and standard deviations in [0.2, 0.45]; (iii) $\mathbf{F}_{b_3}$ contains PDFs with means in [0, 1] and standard deviations in [0.1, 0.5]. Weighting each information equally and combining yields the envelopes shown in (iv).

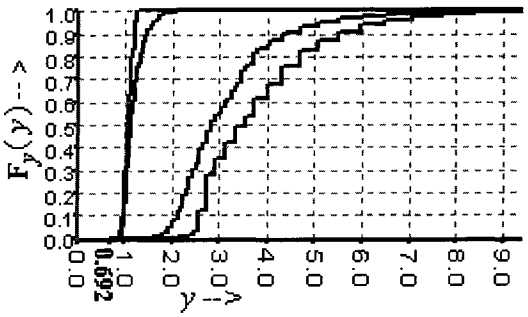



Figure 8. Results for Problem 5a when *a* and *b* are independent of each other (left and right nested envelopes) and when their dependency is unknown (left and right enclosing envelopes). Information about *a* is the three independent, equally weighted, nested intervals [0.5, 0.7], [0.3, 0.8], and [0.1, 1]; *b* is as in Figure 7 (iv).

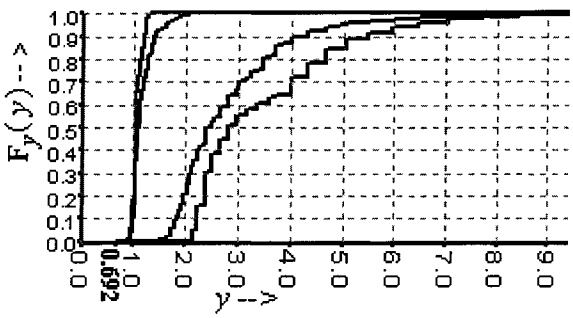



Figure 9. Results for Problem 5b when *a* and *b* are independent (nested envelopes) and when their dependency is unknown (enclosing envelopes). Information about *a* is three independent, equally weighted, overlapping intervals. Information about *b* is similar to that given in Problem 5a, except that the intervals for the means and standard deviations are different [23].

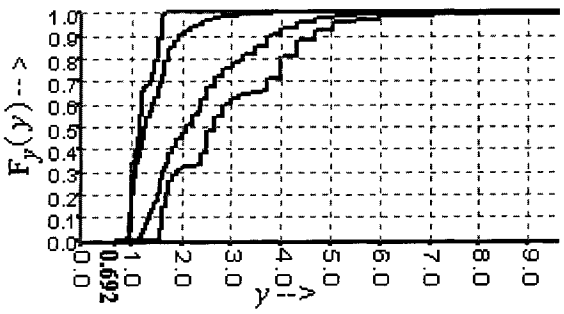



Figure 10. Results for Problem 5c when *a* and *b* are independent (nested envelopes) and when their dependency is unknown (enclosing envelopes). Information about *a* and *b* are similar in format to that given in Problems 5a and 5b.

## Problem 6: an interval for *a* and a distribution for *b*

When *b* is represented using envelopes, as in Challenge Problem 4, the envelopes may be converted into a set of intervals with associated probability masses and the solution obtained as described in Section 2.3. The salient difference in Problem 6 is that *b* is a single distribution rather than a family of possible distributions. To solve this, we enclosed *b* with left and right staircase-shaped envelopes, then convert that pair of envelopes into intervals and their probabilities, and finally solved as before. Figure 11 shows the discretization we used for *b*. The southeast corners of the light left envelope touch northwest corners of the dark right envelope; *b* passes through those contact points. The path of *b* is not fully defined by the discretization over the rectangular regions between contact points, but is constrained to stay between the envelopes. Finer discretizations will constrain the path more, using more contact points and smaller rectangles between contact points. Such a discretization is safe in that it encloses rather than approximating *b*. Figure 11 also shows the result, *y*.

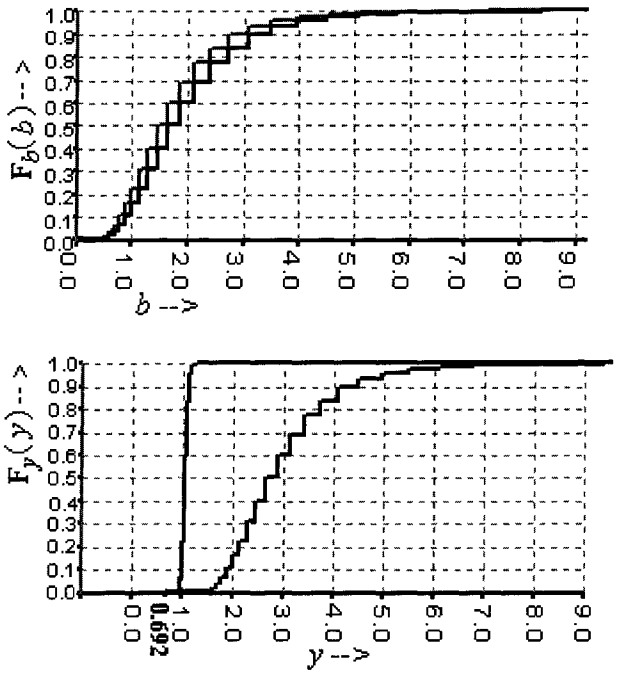



Figure 11. Discretization for $b$ in Problem 6, top, and the resulting envelopes for result $y=(a+b)^a$, bottom, where $b$ is given as a lognormal PDF with mean 0.5 and standard deviation 0.5. Information about $a$ is given as the equally credible intervals [0.1, 0.4], [0.5, 0.7], and [0.8, 1].

## Problem B: the spring system

This problem involves calculating $D_s$ from the equation $D_s = \dfrac{k}{\sqrt{\left(k - m\omega^2\right)^2 + (c\omega)^2}}$.

Note the presence of four variables on the RHS. Each can be converted into a set of intervals and associated probabilities, the form that DEnv requires for marginals, as follows.

(1) For $c$, a set of equally weighted intervals is given. This may be converted into a set of intervals and probabilities as described for Problem 2 (Section 2.1).

(2) For $\omega$, a single interval-parameterized family of PDFs is given. This may be converted into a set of intervals and probabilities as described for Problem 4 (Section 2.3).

(3) For $m$, a distribution is given. This may be converted to a set of intervals and probabilities as described for Problem 6 (Section 2.5).

(4) For $k$, three equally credible, interval-parameterized families of PDFs are given. This set may be converted into a set of envelope pairs, these combined into one pair, and that pair converted into a set of intervals and probabilities, as described for Problem 5 (Section 2.4).

At this point, a joint distribution tableau is needed that generalizes Table 1 to 4 dimensions. It will have one marginal for each of the 4 variables. The interval for each interior cell is determined by the intervals of its four corresponding marginal cells. Thus the interval in the interior cell associated with probability mass $p_{wxyz}$ is the range possible for $D_s$ if $c \in c_w, \omega \in \omega_x, m \in m_y$, and $k \in k_z$. Interval methods can ensure that the computed interior cell intervals are not too narrow and, if too wide, are only slightly so. The four variables are given as independent, and using the standard statistical definition of independence that we have been using, the probability mass of each interior cell is the product of the masses of its four corresponding marginal cells. Once the interior cells are filled in with their intervals and probabilities, the set of interior cells may be used to generate envelopes around $D_s$ by applying a 4-D generalization of Equations (1) or, equivalently, by numbering them consecutively and applying Equations (3). The Statool software applied to the $(a+b)^a$ problems herein (Berleant et al. 2003 [3]) does not currently handle 4-D tableaus but an ad hoc program was written to compute the answer (Figure 12).

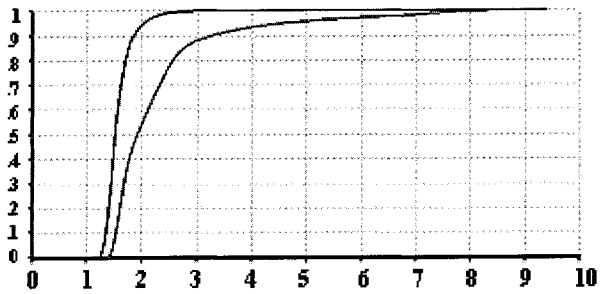



Figure 12. Envelopes around the CDF of $D_s$ in the spring system (Challenge Problem B).

## Combining information

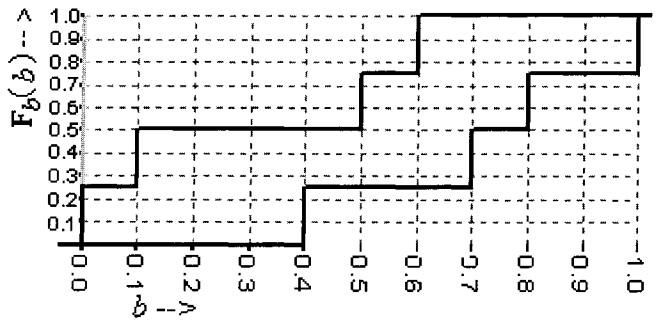
Some of the challenge problems specify $n$ information sources of equal credibility. An important ambiguity is in the likelihood that none of the information sources are correct. This is discussed next. Then in section 3.1 we discuss information equivalence, the fact that different sets of intervals and their probabilities offered as information can be equivalent in significant ways.

### Ambiguity in the likelihood that no source of information is correct

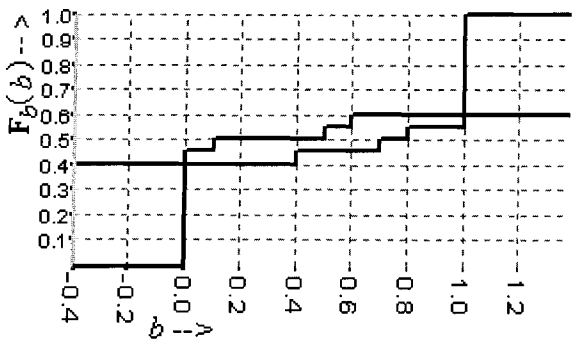
Although multiple information sources are given as having equal credibility in the challenge problems, the credibility of the true value being inconsistent with all of them (call this the *in*credibility) is unspecified. At one extreme the true value might be guaranteed to be consistent with at least one information source (zero incredibility). At the other extreme the credibilities of the information sources, though equal, might be negligible.

This ambiguity has serious implications. For example, envelopes around the CDF of $b$ in Problem 2c can differ dramatically under different resolutions of this ambiguity. The envelopes in Figure 13 (top) were obtained under the assumption of zero incredibility, meaning the actual value of $b$ must be binned in one of the four intervals provided by the four information sources, leading to a probability assignment of 0.25 to each. The envelopes in

Figure 13 (bottom) were obtained from those in Figure 13 (top) by assigning the probability 0.05 instead of 0.25 to each bin. The four interval-valued bins then have a collective probability of (4)·(0.05)=0.2, implying an incredibility of 0.8. If we assume an incredibility of 0.8 and split it evenly into a 0.4 probability that the value of $b$ is below all of the given intervals and a 0.4 probability that it is above all of them, the envelope pair shown will result. It consists of left and right envelopes that touch (without crossing) at two points. The middle portion, between the contact points, is like Figure 13 (top), except scaled and shifted to start at 0.4 on the vertical axis and end at 0.6. It should be noted that solutions given earlier to challenge problems assumed at least one source of information is correct (i.e. that there is no incredibility).




**Envelopes around the cumulation for $b$ in Problem 2c, assuming no incredibility.**




***p*(all items of information are below the actual value)=0.4 and**

***p*(all items of information are above the actual value)=0.4**

Figure 13. Envelopes around the cumulation for $b$ in Problem 2c under two of the infinite number of possible assumptions about the probability that the actual value is not in any of the intervals given as information.

Another example of the effects of considering the possibility that no source of information is correct is that alternative solutions to Problems 3a-3c become plausible, and even arguably more plausible than the solutions given earlier. In Problems 3a-3c the three sources of information about $a$ and four sources about $b$ are stated to be all equally credible. This suggests that the collective credibility of all the information about $a$ is less than the collective credibility of all the information about $b$, since there are fewer sources of information abut $a$. Consequently it would make sense to model these problems with a credibility of ¼ for each of the 3 sources of information about $a$, the same as for each of the 4 sources of information about $b$. Then a fourth possibility for $a$, that its value is given by some unstated or unknown information, will also have a credibility of ¼. This meets the requirement that all 7 information sources have equal credibility. If the ¼ incredibility for $a$ is spread over, for example, [1, 1000] then modeling credibility with probability implies that the CDF of $a$ is ¾ at $a$=1 because of the given information sources, rising to 1 at $a$=1000. Consequently the CDF of $y=(a+b)^a$ would reach 1 only at $y=(1000+1)^{1000}$.

Because all of the challenge problems involving $(a+b)^a$ specified either [0.1, 1] or a subset thereof as information about $a$, we used that interval instead of [0, 1000] as the domain of $a$, and explored the implications of a ¼ incredibility for $a$ in the context of Problem 3c. This incredibility was divided equally between the two intervals within the [0.1, 1] domain that were not covered by any of the three intervals given by the three information sources. This means that intervals for $a$ received the following probability assignments:

$p(a\in[0.1,0.4])=0.25$

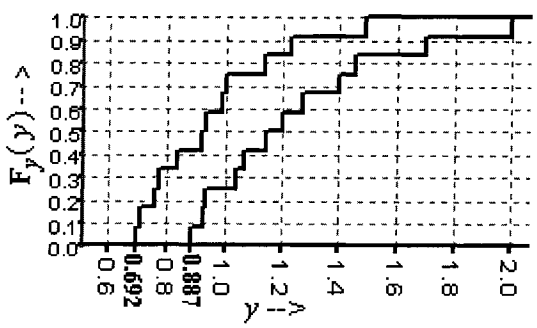
$p(a\in(0.4,0.5))=0.125$

$p(a\in[0.5,0.7])=0.25$

$p(a\in(0.7,0.8))=0.125$

$p(a\in[0.8,1.0])=0.25$ (3)

(where square brackets designate closed intervals and round brackets designate open intervals, although whether any interval is open or closed does not affect the conclusions in this example). Figure 14 (top) shows the resulting envelopes for $y=(a+b)^a$ under the no incredibility condition, while Figure 14 (bottom) shows envelopes obtained under the incredibility scenario of Equations (3). Numerous differences exist between the two results. One is that the heights of the right envelopes differ noticeably at $y=1.8$. Another is that under the incredibility scenario of Equations (3) it is possible that $a\in[0.4, 0.5]$ and $b\in[0, 0.2]$, in which case $y$ can be no higher than 0.837, so the right envelope rises at 0.837. In contrast the right envelope under the no incredibility scenario rises starting at a higher number, 0.887.




**No incredibility in information about *a*.**

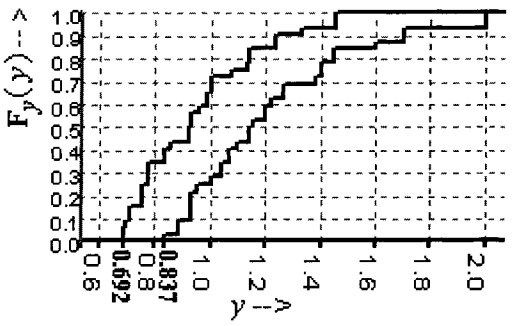



Figure 14. Solutions to Problem 3c under two interpretations of information about *a*. In both, *a* and *b* were assumed independent.

## Information equivalence

***Lemma* 1.** It is possible for two different sets of intervals and associated probabilities to have identical envelopes.

***Proof.*** By example (Figure 15). □

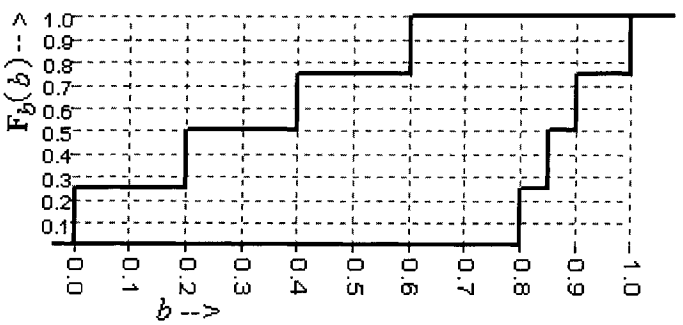



Figure 15. Envelopes around the cumulation of *b* in Problem 2a. They can be generated by Equations (3) from either the pieces of information given in [23], namely [0.6, 0.8], [0.4, 0.85], [0.2, 0.9], and [0, 1], or alternatively from the pieces of information [0, 0.8], [0.2, 0.85], [0.4, 0.9], and [0.6, 1].

***Lemma* 2.** Consider two values $c$ and $d$, each described with a different set of interval-valued pieces of information such that the envelopes computed by Equations (3) around the CDF for $c$ are identical to those computed for $d$. Then, given function $g$ and information about value $e$, the best-possible enclosing envelopes around the CDF of value $g(c,e)$ are *not* necessarily identical to those around the CDF of value $g(d,e)$.

***Proof.*** We will find a unary function $g_1(.)$ for which the envelopes around the CDF of value $g_1(c)$ are not the same as the envelopes for $g_1(d)$. Since a unary function $g_1(.)$ is convertible to an equivalent binary function $g(.,.)$ for which the second argument either does not affect its value or does not affect its value significantly, the lemma will be proved.

Consider function $g_1(.)$ shown in Figure 16. Let the CDF of $c$ be partially defined by intervals [1, 4] and [2, 3], each with probability 0.5. Let the CDF of $d$ also be partially defined, in this case with the intervals [1, 3] and [2, 4], each with probability 0.5. The envelopes around $c$ are identical to those around $d$. This can be seen by applying Equations (3), which gives the same envelopes in both cases because both sets of information have the same low bounds and associated probabilities, and the same high bounds and associated probabilities. Although which low bound is in the same interval as which high bound differs for $c$ and $d$, this does not affect the calculations of Equations (3).

Although the envelopes for $c$ and $d$ are identical, the envelopes for $g_1(c)$ and $g_1(d)$ are *not* identical. Figure 16 shows that $\mathbf{g}_1([1,3])=[6,8] \neq [6,7]=\mathbf{g}_1([2,3])$ for interval extension $\mathbf{g}_1(.)$ of real function $g_1(.)$, while $\mathbf{g}_1([1,4])=[6,9]=\mathbf{g}_1([2,4])$. Consequently applying Equations (3) to the information about $g_1(c)$, intervals [6, 7] and [6, 9] each with probability 0.5, gives a different pair of envelopes than applying Equations (3) to the information about $g_1(d)$, intervals [6, 8] and [6, 9] each with probability 0.5. Readers may enjoy verifying this for themselves.

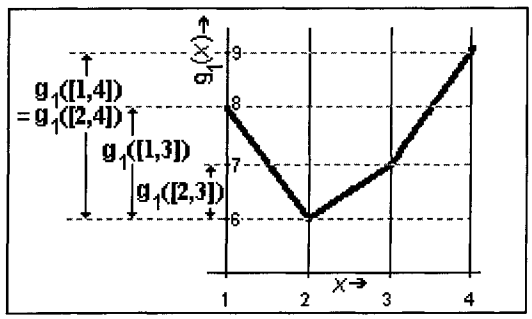



Figure 16. A function $g_1(x)$ and projections of 4 intervals for $x$ from the horizontal to the vertical axis. For example $g_1(x)$ is within the range [6, 8] when $x$ is within [1, 3] (with the 6 occurring when $x$=2).

So far this section has shown that two uncertain values described by different sets of intervals and their probabilities can have identical envelopes. A function was found that takes different inputs with identical envelopes, and gives outputs with differing envelopes. On the other hand some other functions, when given different inputs with identical envelopes, will be guaranteed to output identical envelopes. Although not a full characterization of such envelope-preserving functions, the following holds.

***Theorem* 1: envelope invariance for monotonic functions.** Let the available information about value $c$ be described by a set of intervals and their associated probability masses, and similarly for $d$. Then Equations (3) may be applied to get envelopes around the CDFs for $c$ and for $d$. If

(i) the envelopes for $c$ are identical to those for $d$,

(ii) the set of intervals and associated probability masses for $c$ differs from the set for $d$,

(iii) $g(.,.)$ is increasing in both arguments, and

(iv) $c$ and $d$ are statistically independent in the usual sense,

then the envelopes around the CDF for value $g(c, .)$ are identical to the envelopes around the CDF for value $g(d, .)$.

The strategy for establishing the theorem is to dissociate the effects on the result envelopes of the interval low bounds in the marginals from the effects of the high bounds. Then for constructing result envelopes it doesn't matter which marginal interval low bounds are in the same interval with which marginal interval high bounds, so that different information can lead to the same result envelopes.

- First consider point (i). Since $c$ and $d$ have identical envelopes, each vertical line segment in the left envelope of $c$ has the same location on the horizontal axis and the same bottom-to-top length as a vertical line segment in the left envelope of $d$, and vice versa. Equations (3) imply that the horizontal axis location of each vertical line segment in the left envelope of $c$ equals the low bound of interval(s) in the discretization of $c$, and the pooled mass associated with the interval(s) is the length of the segment. The same is true of the left envelope for $d$, and also for the right envelopes of both $c$ and $d$ except with reference to the interval high bounds instead of their low bounds.

- Next consider point (ii). This can be the case (by Lemma 1).

- Now consider point (iii). If $g(x,y)$ increases monotonically in $x$ and $y$, then the range of $g(x,y)$ over the rectangle $[x_l, x_h]\mathrm{x}[y_l, y_h]$ is $[g(x_l, y_l), g(x_h, y_h)]$. This is because $g(x,y)$, being monotonically increasing, has no local minima or maxima. Thus its minimum and maximum over a rectangle are at the southwest and northeast corners respectively.

  Each interior cell in a joint distribution tableau is determined by two marginal cells (Table 1) whose intervals define a rectangle. Therefore the interval in each interior

cell has its low bound determined by the low bounds of its corresponding marginal cell intervals. Consequently the locations on the horizontal axis of the vertical line segments of the left envelope of $g(x,y)$, because they are determined by the low bounds of the interior cell intervals per Equations (1), are ultimately derived from the low bounds of the marginal intervals. The situation is analogous for high bounds and the right envelope. Therefore which marginal interval low bounds occur in the same interval with which high bounds is irrelevant in determining the *horizontal axis locations of vertical line segments in the result envelopes.*

- Finally, consider point (iv). In the case of independence, the probability mass of each interior cell in a joint distribution tableau is the product of the masses of its two corresponding marginal cells, and the intervals associated with these masses have no effect. Thus the interval high bounds do not directly affect the mass computations when computing the left envelope. Similarly, interval low bounds do not directly affect the mass computations for the right envelope. It is these mass computations that determine the lengths of the vertical line segments in the envelopes. Therefore which marginal interval low bounds occur in the same interval with which high bounds is irrelevant in determining the *lengths of vertical line segments in the result envelopes.*

Thus neither the horizontal axis locations of the vertical line segments in the left envelope nor their lengths are affected by any interval high bounds in the tableau, only low bounds. Similarly, neither the horizontal axis locations of the vertical line segments in the right envelope nor their lengths are affected by any interval low bounds in the tableau, only high bounds. Therefore which marginal interval high bounds occur in the same intervals with

which marginal interval low bounds does not affect the envelopes and the theorem is established.

*Further comments.* The irrelevance of which high bound is associated with which low bound explains why the envelopes of Figure 15 are implied by both of the two different information sets noted in that figure.

Further work is needed to better understand equivalence of information sets in the presence of unknown dependency and correlation constraints, as well as in cases where $g(x,y)$ is not monotonically increasing in $x$ and $y$. Some recent work relevant to such questions includes Hall and Lawry (this issue [15]), Ferson and Kreinovich [12], and Kreinovich et al. 2001 [18].

## Conclusion

The challenge problems of Oberkampf et al. [23] provide a valuable opportunity to compare techniques and their implementations for computing functions of random variables whose samples are expressed using intervals, distributions, or families of distributions. Monte Carlo techniques form a conceptually graspable class of algorithms but give results complicated by random noise. Probabilistic Arithmetic, random set theory, and joint distribution tableaus, which can be manipulated by the DEnv algorithm, are the most widely reported alternatives. Of these three, joint distribution tableaus do not require an understanding of copulas or random sets, so are arguably easier to understand. DEnv can compute functions of random variables when they are either: (*i*) independent of each other, (*ii*) have some other specific dependency relationship, (*iii*) have an unknown dependency relationship, or (*iv*) have a dependency that is partially characterized. Results provided by DEnv are consistent with those provided by the other techniques to date. Additional work is needed to achieve further advances in such directions as more flexible handling of partially

characterized dependency. Ultimately, crossing the bridge from 2nd order probabilistic results to decisions is likely to be a key factor in their achieving widespread use.

## Acknowledgements

The authors are grateful to Lizhi Xie for building an initial version of Statool suitable for continued development, to Gerald Sheblé for motivating advances in Statool through extensive discussions on the needs of applications in the electric power industry, especially power systems economics, and to Scott Ferson and Vladik Kreinovich for valuable discussions and well-timed encouragement. We also acknowledge support from PSERC to G. Sheblé, D. Berleant, and R. Thomas.

## References

[1] Berleant D. Automatically verified reasoning with both intervals and probability density functions. *Interval Computations* (1993 No. 2), pp. 48-70.

[2] Berleant D and Goodman-Strauss C. Bounding the results of arithmetic operations on random variables of unknown dependency using intervals. *Reliable Computing* 4 (2) (1998), pp. 147-165.

[3] Berleant D, Xie L, and Zhang J. Statool: a tool for Distribution Envelope Determination (DEnv), an interval-based algorithm for arithmetic on random variables. *Reliable Computing* 9 (2) (2003), pp. 91-108.

[4] Berleant D, Zhang J, Hu R, and Sheblé G. Economic dispatch: applying the interval-based distribution envelope algorithm to an electric power problem. *SIAM Workshop on Validated Computing 2002 (Extended Abstracts)*, Toronto, May 23-25, 2002, pp. 26-31.

[5] Berleant D and Zhang J. Using correlation to improve envelopes around derived distributions. *Reliable Computing*, accepted. www.public.iastate.edu/~berleant.

[6] Colombo AG and Jaarsma RJ. A powerful numerical method to combine random variables. *IEEE Transactions on Reliability* R-29 (2) (1980), pp. 126-129.

[7] Couso I, Moral S, and Walley P. Examples of independence for imprecise probabilities, *1st Int. Symp. On Imprecise Probabilities and their Applications*, Ghent, Belgium, 1999. decsai.ugr.es/~smc/isipta99/proc/proceedings.html.

[8] Ferson S. *RAMAS Risk Calc 4.0: risk assessment with uncertain numbers*. Lewis Press, Boca Raton, 2002.

[9] Ferson S. What Monte Carlo methods cannot do. *Journal of Human and Ecological Risk Assessment* 2 (4) (1996), pp. 990-1007.

[10] Ferson S, Ginzburg L, and Akçakaya R. Whereof one cannot speak: when input distributions are unknown, *Risk Analysis,* to appear.

[11] Ferson S and Hajagos J. Rigorous and (often) best-possible answers to arithmetic problems. *Reliability Engineering and System Safety*, this issue.

[12] Ferson S and Kreinovich V. Representation, elicitation, and aggregation of uncertainty in risk analysis – from traditional probabilistic techniques to more general, more realistic approaches: a survey. Manuscript, vladik@cs.utep.edu.

[13] Fetz T and Oberguggenberger M. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety,* this issue.

[14] Frank MJ, Nelsen RB, and Schweizer B. Best-possible bounds for the distribution of a sum – a problem of Kolmogorov. *Probability Theory and Related Fields* 74 (1987), pp. 199-211.

[15] Hall J and J Lawry. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. *Reliability Engineering and System Safety*, this issue.

[16] Ingram GE, Welker EL, and Herrmann CR. Designing for reliability based on probabilistic modeling using remote access computer systems. *Proc. 7th Reliability and*

*Maintainability Conference*, American Society of Mechanical Engineers, 1968, pp. 492-500.

[17] Kaplan S. On the method of discrete probability distributions in risk and reliability calculations, applications to seismic risk assessment. *Risk Analysis* 1 (3) (1981), pp. 189-196.

[18] Kreinovich V, Langrand C, and Nguyen HT. Combining fuzzy and probabilistic knowledge using belief functions. *Proc. 2nd Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications 2001*, Hanoi, Dec. 7-8, pp. 191-198.

[19] Lodwick W and Jamison KD. Estimating and validating the cumulative distribution of a function of random variables: toward the development of distribution arithmetic. *Reliable Computing* 9 (2) (2003), pp. 127-141.

[20] Moore RE. Risk analysis without Monte Carlo methods. *Freiburger Intervall-Berichte*, 84/1, 1984, pp. 1-48.

[21] Nelsen RB. *An Introduction to Copulas.* Lecture Notes in Statistics, Vol. 139, 1999, Springer-Verlag.

[22] Neumaier A. Clouds, fuzzy sets and probability intervals. Submitted. www.mat.univie.ac.at/~neum/ms/papers.html.

[23] Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewics SF, and Ferson S. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety*, this issue.

[24] Red-Horse J. and Benjamin AS. A probabilistic approach to uncertainty quantification with limited information. Reliability Engineering and System Safety, this issue.

[25] Regan H, Ferson S, and Berleant D. Equivalence of five methods for bounding uncertainty. *Int. Journal of Approximate Reasoning*, accepted pending revisions. Draft at www.public.iastate.edu/~berleant.

[26] Sandia National Laboratory. *Epistemic Uncertainty Workshop*, August 6-7, 2002, Albuquerque. Presentations and papers are at www.sandia.gov/epistemic/.

[27] Sheblé G and Berleant D. Bounding the composite value at risk for energy service company operation with DEnv, an interval-based algorithm. *SIAM Workshop on Validated Computing 2002 (Extended Abstracts)*, Toronto, May 23-25, 2002, pp. 166-171.

[28] Springer MD. *The Algebra of Random Variables*, John Wiley and Sons, New York, 1979.

[29] Tonon F. Using random set theory to propagate epistemic uncertainty through a mechanical system. *Reliability Engineering and System Safety*, this issue.

[30] Williamson RC and Downs T. Probabilistic Arithmetic I: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4 (1990), pp. 89-158.

# CHAPTER 6. GENERAL CONCLUSION

This dissertation has described various ways to improve the results of arithmetic operation on random variables by using partial information about dependency. Interval analysis provides the fundamental method to do the computations.

When calculating a derived random variable, the dependency of the marginals is an important issue that affects results. Generally the exact joint distribution may be unknown, while the marginal distributions are given. Pearson correlation is a linear measure of dependency. This correlation can be used to improve the results even if this information is not exact described but is given as a range. The algorithms are given in this work.

Parametric uncertainty often happens. For this situation, the CDF is not single curve, but a family whose members correspond to different parameter values. We investigated some common distributions. The closed form expressions for CDFs were used with interval parameters. Then these bounded CDF families were used as inputs to algorithms that manipulate distributions and bounded spaces defining families of distributions.

Another important potential source of partial information concerns the joint distribution and the result random variable. We focus on the following kinds of partial information:

- Knowledge about probabilities of specified areas of the joint distribution.
- Knowledge about probabilities of specified ranges of values of the derived variable.
- Known relationships among the probabilities of different areas of the joint distribution.
- Known relationships among the probabilities of different ranges of the derived random variable.

For each situation, algorithmic modifications to handle it are given. The results are improved envelopes that are closer than they would be without this information.

Uncertainties exist widely in realistic models. We present solutions to sample models proposed by Sandia National Laboratory. Also these methods can be used in different fields, such as decision analysis, Pert networks, and computer engineering.

Processing uncertainty with interval based methods has been demonstrated to be an effective method. DEnv algorithm is a method that uses intervals. There are numerous opportunities for future research on using partial information about dependency and on applying DEnv and related techniques in practical situations.

# APPENDIX. STATOOL SOFTWARE

Statool is a useful tool to do uncertainty operations. New functions were added. Also the usability was improved. The visualization of the results has also been significantly improved.

## Algorithms implemented in Statool

In this part, detailed information about narrowing the CDF envelops in the "correlation setting" pop up window is given. See the following Figure. This window is obtained by clicking on the radio button labelled "Known dep." in the main screen of Statool. We list the approaches which are used to calculate the theoretical correlation and expectation of XY. Also we demonstrate how to get constraints for linear programming from the different setting values.

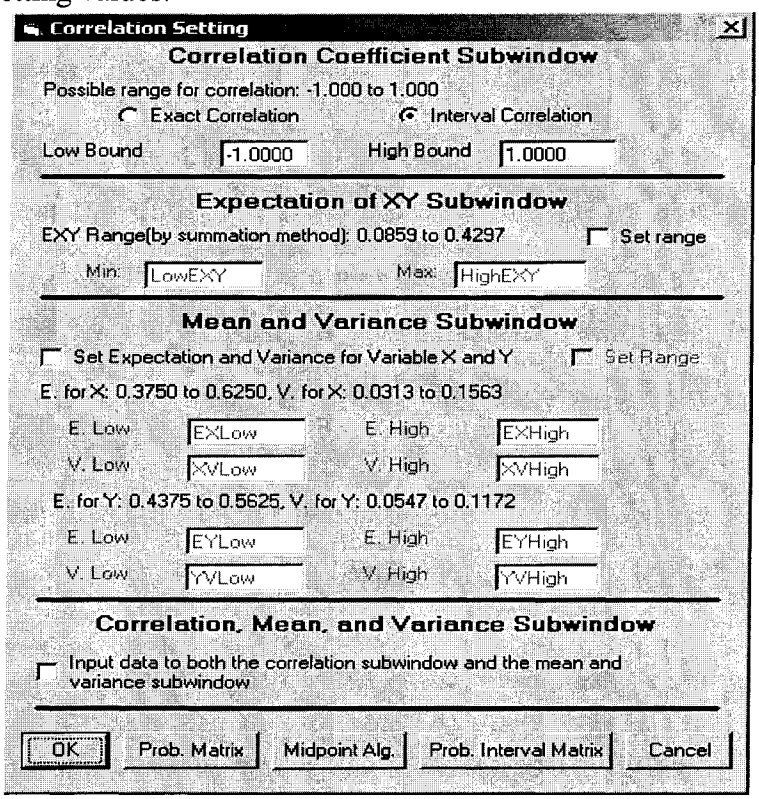



Figure 1. "Correlation Setting" popup window.

## Obtaining expectation of *XY* based on the join distribution: $E_t$

This section illustrates how to figure out the possible range of the expectation of *XY* if the marginal distributions of *X* and *Y* are known.

Assume that the marginal distributions of *X* and *Y* are known, as listed in the following table.

Table 1. The joint distribution tableau.

| Y ↓ X→ | $X_1=[x_{1l},x_{1h}]$ | ... | $X_n=[x_{nl},x_{nh}]$ | |
|---|---|---|---|---|
| $Y_1=[y_{1l},y_{1h}]$ | $p_{11}$ | ... | $p_{1n}$ | $p_{y1}$ |
| ... | ... | ... | ... | ... |
| $Y_m=[y_{ml},y_{mh}]$ | $p_{m1}$ | ... | $p_{mn}$ | $p_{ym}$ |
| | $p_{x1}$ | ... | $p_{xn}$ | 1 |

According to the definition of the expectation of $X*Y$, $EXY$, we have

$EXY=\sum_{i=1}^{n}\sum_{j=1}^{m}X_i*Y_j*p_{ij}$. Here *Xi* and *Yj* are interval values. Based on interval multiplication,

$$X_i*Y_j*p_{ij}=[x_{il},x_{ih}]*[y_{jl},y_{jh}]*p_{ij}$$
$$=[\min(x_{il}*y_{jl},x_{il}*y_{jh},x_{ih}*y_{jl},x_{ih}*y_{jh}),\max(x_{il}*y_{jl},x_{il}*y_{jh},x_{ih}*y_{jl},x_{ih}*y_{jh})]*p_{ij}$$

Let $MinXY_{ij}=\min(x_{il}*y_{jl},x_{il}*y_{jh},x_{ih}*y_{jl},x_{ih}*y_{jh})$ and

$MaxXY_{ij}=\max(x_{il}*y_{jl},x_{il}*y_{jh},x_{ih}*y_{jl},x_{ih}*y_{jh})$

Then $EXY=\sum_{i=1}^{n}\sum_{j=1}^{m}[MinXY_{ij},MaxXY_{ij}]*p_{ij}$

There also are the constraints on the $p_{ij}$'s from the marginal distributions. These are the row and column constraints, as follows;

$\sum_{i=1}^{m}p_{ij}=p_{xj}$ for $j$=1 to $n$

$\sum_{j=1}^{n}p_{ij}=p_{yi}$ for $i$=1 to $m$

Here, only the $p_{ij}$, $i$=1 to $m$, $j$=1 to $n$ are unknown. Our objective is to find the minimum and maximum values possible for *EXY*. Since each $p_{ij}$ is non-negative, the minimum value of *EXY* is obtained by minimizing $\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij}$ and the maximum value of *EXY* is obtained by maximizing $\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij}$. Therefore two linear programs are constructed to get the minimum and maximum values of *EXY*.

Minimum value:

Minimize $\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij}$

Subject to:

$\sum_{i=1}^{m} p_{ij} = p_{xj}$ for $j$=1 to $n$

$\sum_{j=1} p_{ij} = p_{yi}$ for $i$=1 to $m$

Maximum value:

Maximize $\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij}$

Subject to:

$\sum_{i=1}^{m} p_{ij} = p_{xj}$ for $j$=1 to $n$

$\sum_{j=1} p_{ij} = p_{yi}$ for $i$=1 to $m$

After solving these two linear programming problems, the minimum and maximum values of *EXY* are obtained and are recorded as *Emin* and *Emax*. These values are presented in the "Expectation of *XY* sub-window" of the "Correlation Setting" pop up window.

## Theoretical correlation

Although the marginal distributions don't determine the exact correlation between two random variables, they often constrain it to some extent. In the following, we will show how to compute the possible correlation range from the marginal distributions.

From the definition of correlation, $\rho = \frac{EXY - EX * EY}{\sqrt{Var(X) * Var(Y)}}$ where Var(X) and Var(Y) are the variances of $X$ and $Y$. Rearranging, $EXY = EX * EY + \rho\sqrt{Var(X) * Var(Y)}$. From the previous section, the theoretical range of *EXY* from the definition of *EXY* is from *Emin* to *Emax*. Here we have another formula of *EXY* from the definition of correlation. We consider computing the possible range of *EXY* from this new definition. *EXY* can be written as

$$EXY = \left(\sum_{i=1}^{n} X_i p_{xi}\right) * \left(\sum_{j=1}^{n} Y_j p_{yj}\right) + \rho\sqrt{\left(EX^2 - (EX)^2\right) * \left(EY^2 - (EY)^2\right)}$$

$$= \left(\sum_{i=1}^{n} X_i p_{xi}\right) * \left(\sum_{j=1}^{n} Y_j p_{yj}\right) + \rho\sqrt{\left(\sum_{i=1}^{n} X_i^2 p_{xi} - (\sum_{i=1}^{n} X_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} Y_j^2 p_{yj} - (\sum_{j=1}^{m} X_j p_{yj})^2\right)}$$

We define the function

$$F(X,Y) = \left(\sum_{i=1}^{n} X_i p_{xi}\right) * \left(\sum_{j=1}^{n} Y_j p_{yj}\right) + \rho\sqrt{\left(\sum_{i=1}^{n} X_i^2 p_{xi} - (\sum_{i=1}^{n} X_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} Y_j^2 p_{yj} - (\sum_{j=1}^{m} Y_j p_{yj})^2\right)}$$

This is an interval-valued function. We write the corresponding real function as

$$F(x,y) = \left(\sum_{i=1}^{n} x_i p_{xi}\right) * \left(\sum_{j=1}^{n} y_j p_{yj}\right) + \rho\sqrt{\left(\sum_{i=1}^{n} x_i^2 p_{xi} - (\sum_{i=1}^{n} x_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} y_j^2 p_{yj} - (\sum_{j=1}^{m} y_j p_{yj})^2\right)}$$

where $x_i \in X_i$ $i$=1 to $n$, $y_j \in Y_j$ $j$=1 to $m$, and $\rho \in [-1,1]$. In this function, there are $n$+$m$+1 variables and every variable is restricted to the specified interval range. We can use an optimization method to find the minimum and maximum value for *F(x,y)* and record them as *Fmin* and *Fmax*. (This is a nonlinear optimization problem).

Now we get two ranges for *EXY* from the different formulas. Since both are true, we exploit both by intersecting them. Call the low and high bounds the intersection *Gmin* and

$Gmax$. Then $\rho \geq \dfrac{G\min - \left(\sum_{i=1}^{n} x_i p_{xi}\right) * \left(\sum_{j=1}^{n} y_j p_{yj}\right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^{\,2} p_{xi} - (\sum_{i=1}^{n} x_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} y_j^{\,2} p_{yj} - (\sum_{j=1}^{m} y_j p_{yj})^2\right)}}$

and $\rho \leq \dfrac{G\max - \left(\sum_{i=1}^{n} x_i p_{xi}\right) * \left(\sum_{j=1}^{n} y_j p_{yj}\right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^{\,2} p_{xi} - (\sum_{i=1}^{n} x_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} y_j^{\,2} p_{yj} - (\sum_{j=1}^{m} y_j p_{yj})^2\right)}}$.

The values of $x_i$ and $y_j$ used to compute *Fmin* and *Fmax* are used again here to compute the bounds on $\rho$.

Since we just want to get a safe range for correlation, not necessarily the narrowest possible range, we are done.

A more accurate range for correlation could be gotten directly from computing the min and max of $\rho = \dfrac{EXY - EX * EY}{\sqrt{Var(X) * Var(Y)}}$. This is a complex nonlinear optimal problem. The range is presented in the "Correlation Coefficient Subwindow" of the "Correlation Setting" popup window.

## Mean and variance

Theoretical ranges for mean and variance of a discretized operand are calculated by the program. These values are directly obtained according to the definitions.

By definition, the expectation of random variable $X$ is $EX = \sum_{i=1}^{n} X_i p_{xi}$. Since *Xi* is an interval value, $EX = \sum_{i=1}^{n} [x_{il}, x_{ih}] * p_{xi} = \left[\sum_{i=1}^{n} x_{il} * p_{xi}, \sum_{i=1}^{n} x_{ih} * p_{xi}\right]$. So the bounds on $EX$ are

obtained. The same method is used to handle operand $Y$. The bounds on $EY$ are $\sum_{j=1}^{m} y_{jl} * p_{yj}$ and $\sum_{j=1}^{m} y_{jh} * p_{yj}$ .

Variances for $X$ and $Y$ are a little more complex to obtain. Based on the definition, the variance of $X$ is $Var(X) = EX^2 - (EX)^2 = \sum_{i=1}^{n} X_i{}^2 p_{xi} - (\sum_{i=1}^{n} X_i p_{xi})^2$. Here each $Xi$ is an interval value. This is a problem of evaluation of an interval function. We define a real function $V(x) = \sum_{i=1}^{n} x_i{}^2 p_{xi} - (\sum_{i=1}^{n} x_i p_{xi})^2$ and each $x_i \in X_i$ $i$=1 to $n$. Since all $p_{xi}$ are known, the optimization method can be adapted to compute the min and max values of function $V(x)$ as *VXmin* and *VXmax*. The similar method is used to variance of $Y$. Let $V(y) = \sum_{j=1}^{m} y_j{}^2 p_{yj} - (\sum_{j=1}^{m} y_j p_{yj})$ and $y_j \in Y_j$ $j$=1 to $m$. Then the bounds on the variance of $Y$ are obtained and recorded as *VYmin* and *VYmax*. These ranges are presented in the "Mean and Variance Subwindow" of the "Correlation Setting" popup window.

## Constraints from setting the range of correlation

In this section, we demonstrate how to get extra constraints if the user sets the range of correlation in the "Correlation Coefficient Subwindow" of the "Correlation Setting" popup window.

From section 2,

$$EXY = \sum_{i=1}^{n}\sum_{j=1}^{m} [MinXY_{ij}, MaxXY_{ij}] * p_{ij} = \left[\left(\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij}\right), \left(\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij}\right)\right],$$

since $p_{ij}$ is non-negative.

From section 3,

$$EXY = F(X,Y)$$
$$= \left(\sum_{i=1}^{n} X_i p_{xi}\right) * \left(\sum_{j=1}^{n} Y_j p_{yj}\right) + \rho \sqrt{\left(\sum_{i=1}^{n} X_i{}^2 p_{xi} - (\sum_{i=1}^{n} X_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} Y_j{}^2 p_{yj} - (\sum_{j=1}^{m} X_j p_{yj})^2\right)}$$

Using the real function

$$F(x,y) = \left(\sum_{i=1}^{n} x_i p_{xi}\right) * \left(\sum_{j=1}^{n} y_j p_{yj}\right) + \rho \sqrt{\left(\sum_{i=1}^{n} x_i{}^2 p_{xi} - (\sum_{i=1}^{n} x_i p_{xi})^2\right) * \left(\sum_{j=1}^{m} y_j{}^2 p_{yj} - (\sum_{j=1}^{m} y_j p_{yj})^2\right)}$$

and $x_i \in X_i$ $i$=1 to $n$, $y_j \in Y_j$ $j$=1 to $m$, and given range for correlation $\rho$, the minimum and maximum values of $F(x,y)$ can be calculated by non-linear optimization as in section 3. Call them *Fmin* and *Fmax*.

Based on Berleant & Zhang, two inequalities are defined:

$\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij} \le F\max$ and $\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij} \ge F\min$. These two inequalities form two extra constraints for linear programming since only the $p_{ij}$'s are unknown.

## Constraints from setting the range of *EXY*

If the user sets "*EXY* range" in the "Expection of *EXY*" subwindow of the "Correlation Setting" popup windows, the values that the user provides, *Fmin* and *Fmax*, are used directly to define 2 constraints:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij} \le F\max$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij} \ge F\min$$

These constraints were justified in the section 5 and in Berleant & Zhang.

## Constraints from setting mean and variance of *X* and *Y*

The user can set mean and /or variance in the "Mean and Variance Subwindow" of the "Correlation Setting" popup window. Consider the formula $EXY = EX * EY + \rho\sqrt{Var(X) * Var(Y)}$. If the means and variances of $X$ and $Y$ are known, the value of $EXY$ can be calculated if correlation is also known. From section 5, the range for correlation is computable. We can use this range of correlation to calculate the range of $EXY$. It is clear that computing $EXY$ is interval, not a real number. Let the low bound of $EXY$ be called Fmin and the high bound be called Fmax. Then, $\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij} \le F\max$, and $\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij} \ge F\min$. These constraints are then used by Statool.

## Constraints from setting correlation, mean and variance of *X* and *Y*

In the some situations, the user may have partial information about both correlation, and either mean, variance or both. Here is how the user can choose values for correlation, and mean and/or variance.

First, the user should click on the checkbox button labelled "Input data in both the correlation subwindow and the mean and variance subwindow" in the "Correlation, Mean, and Variance Subwindow" of the "Correlation Setting" popup window. Then the user can set values in both the "Correlation" and "Mean, Variance" subwindows of the "Correlation Setting" popup window.

In section 7, we describe the situation where mean and/or variance are known. If correlation is also input, we can directly use all three in the formula $EXY = EX * EY + \rho\sqrt{Var(X) * Var(Y)}$ to get the value of $EXY$. If either mean or variance is missing, a default range for it may be obtained as described in section 4. Let the low bound of $EXY$ be called $Fmin$ and the high bound be called $Fmax$. Then, $\sum_{i=1}^{n}\sum_{j=1}^{m} MinXY_{ij} * p_{ij} \le F\max$, and $\sum_{i=1}^{n}\sum_{j=1}^{m} MaxXY_{ij} * p_{ij} \ge F\min$. These extra constraints are then added to the LP calls.

## User-scaled visualization

Graph display is a useful way to show results. It is important to clearly and efficiently display the result.

The scale and font are critical factors in displaying the CDF envelopes. The software can generate the display using default settings for scale and font. The following figure illustrates a case.

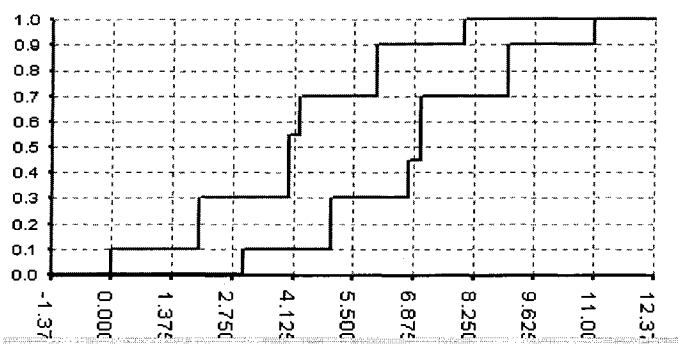



Figure 2. A default result view.

However the default view may not show the clearest possible x-axis scaling. The user can change the scale of x-axis and redraw the CDF envelopes as follows:

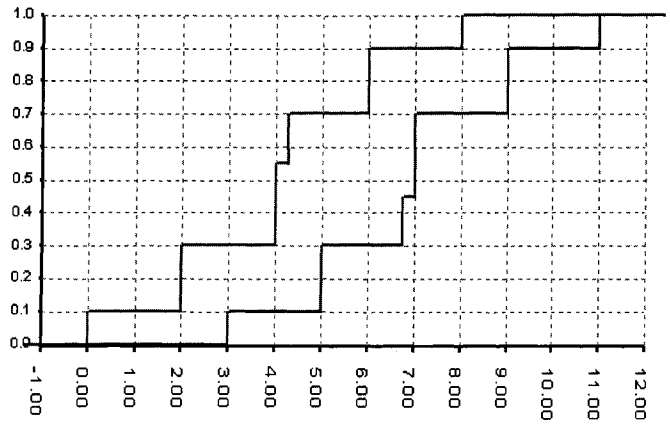



Figure 3. A user-scaled result view.

This way, the user can design the display to match the requirements.

# ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting the research and the writing of this dissertation.

First to my major professor, Dr. Daniel Berleant, I want to thank him for giving me the chance to work with him closely on the research project and for his great advice on my graduate studies in these years. Dr. Berleant read the drafts of my dissertation and gave me a lot of constructive feedback even when he was in vacation with his family. I was always moved when I got his comments and corrections on my thesis via email. I appreciate deeply his time and work. It has been a very good luck for me to have worked with him. What I have learned from him will benefit me in my all life.

I also want to thank Dr. Gerad Sheble with whom I had worked on the research project. Our weekly meetings had been great opportunities for me to learn from him. I have enjoyed the meetings. Also thank you to my committee member, Dr. Lahri, Dr. Wang and Dr. Mao. Thanks also to my friends who give me support and understanding.

I want to say "thank you" to my dear parents for instilling in me the value of endless learning and a thirst for knowledge. Without it, I would not have even been able to achieve this degree.

Finally my heartfelt thanks go to my wife, De, and to my little daughter, Peilan, for their wonderful support, and for not minding the many lost evenings and weekends during my preparation for this dissertation.