# An Interval-Based Tool for Verified Arithmetic on Random Variables of Unknown Dependency

Daniel Berleant and Lizhi Xie
Department of Electrical and Computer Engineering
Iowa State University
Ames, Iowa 50011, USA
berleant@iastate.edu

## Abstract

When random variables possessing arbitrary distribution functions must be added, subtracted, multiplied, etc., Monte Carlo simulation is commonly employed. However, Monte Carlo simulation typically assumes that the distribution functions to be combined are independent, and must assume either independence or some other specific dependency relationship. Discretization of the distribution function followed by a numerical method is another alternative approach. Numerical methods can relax the requirement of Monte Carlo that the distributions must have a known dependency relationship by producing boundary curves within which the cumulative distribution resulting from any dependency relationship must lie. Furthermore, numerical methods can account for discretization error by producing boundary curves that are more inclusive than they would be if the same problem was run with a finer discretization, thereby accounting for discretization error in the results. This paper describes *Statool,* a software tool available in source and binary, that runs an interval-based numerical method for performing arithmetic operations on distribution functions that are either independent or have an unknown dependency relationship, and provides inclusion guarantees deriving from its interval based calculations.

## Introduction

Random variables may be combined using standard operations such as +,-,*, and /. When the random variable operands are assumed independent, results may be calculated using a discretized convolution approach (Ingram et al. 1968; Kaplan 1981). Discretization error may be bounded by an interval based extension (Berleant 1993). When the dependency relationship between the operands is not known, obtaining verified results requires that independence not be assumed, but rather that the entire range of possible dependency relationships be accounted for. While the traditional approach of Monte Carlo simulation does not bound the range of possible results as required for verified computations (Ferson 1996), the desired bounds can be obtained with other techniques. A copula-based approach (Frank et al. 1987) which was significantly extended by Williamson and Downs (1990) has been implemented in a commercially available software system, RiskCalc (Ferson et al. 1998). An interval-based approach is described by Berleant and Goodman-Strauss (1998). Verified operations on random variables of unknown dependency has been applied to problems in risk analysis (e.g. Ferson and Bergman 1995) and we are currently investigating problems in related areas of decision analysis, finance (valuations and bond management), and PERT graphs.

The interval-based and copula-based techniques are compared in Berleant and Goodman-Strauss (1998), where the interval-based technique is explained and shown to have advantages in comparison to the Williamson and Downs (1990) approach that merit its further development.

This paper extends that work by reporting on a software tool, Statool, that may be downloaded and which implements the interval-based technique. Figure 1 shows the startup screen in Statool.

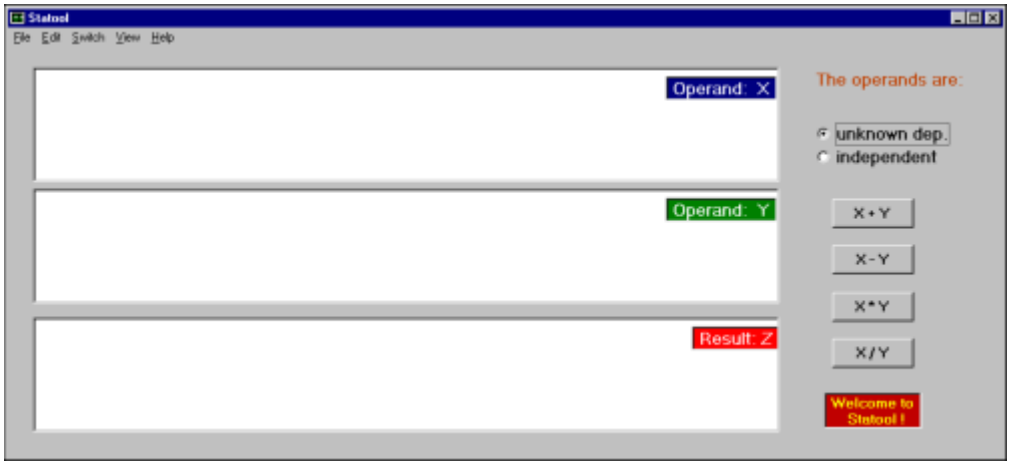


**Figure 1. Statool start screen, consisting of three subwindows. The upper two provide a place to edit or read in operands X and Y, and the lower one provides a place to display the result to be obtained by applying some arithmetic operation to operands X and Y, once they are specified. In the area to the right of the three subwindows, at top, operands X and Y can be declared to be either of unknown dependency, or independent.**

In Statool, probability distribution functions representing the operand random variables can be flexibly edited in histogram form via both a mouse operated GUI (Figure 2) and with type-in boxes that accept numerical input values (Figure 3).

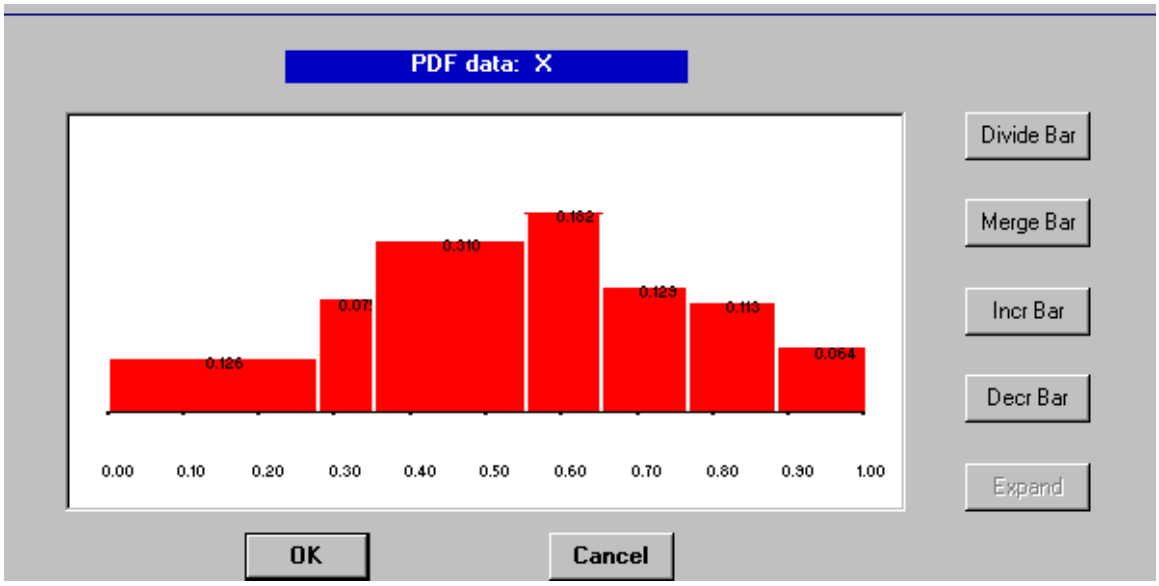


**Figure 2. Editing window for specifying the contents of the top subwindow in Figure 1, labeled "Operand: X." Clicking on "Divide Bar" doubles the number of histogram by dividing each one in half. "Merge Bar" does the opposite. "Incr Bar" and "Decr Bar" change the number of bars by one. A left mouse click above a bar causes the bar to become taller, the amount of increase depending on how far above the bar the click occurs. A left click below the top of the bar causes an analogous shortening of the bar. Similarly, a right click above or below the top of a bar causes the bar to become wider or narrower.**

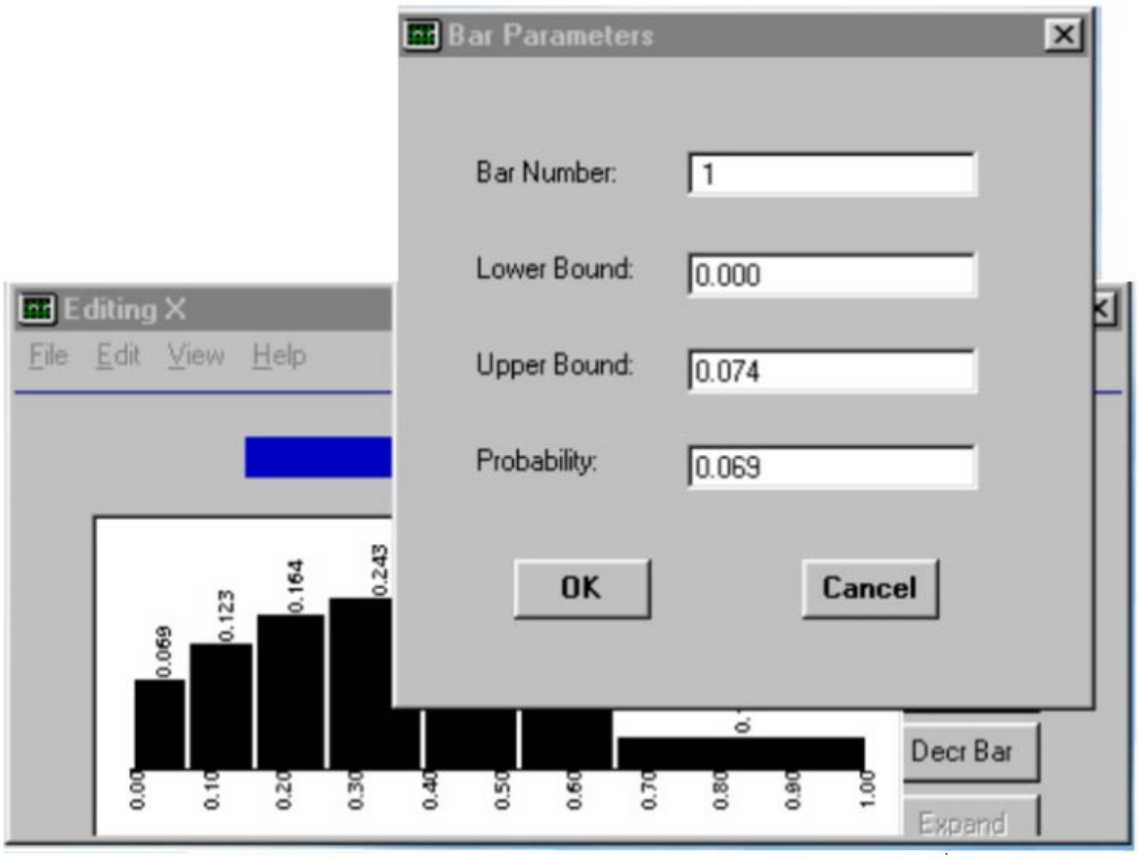


**Figure 3. Numerical inputs can be used to specify characteristics of the discretized PDF by clicking on "Edit." Shown here is a popup window for numerical specification of the 1st bar, initialized to its current state. The entire histogram can also be scaled by specifying the endpoints of the histogram, which default to 0 and 1 as shown.**

Operations currently available are +,-,*, and / (Figure 1). These operations may be invoked either with or without assuming independence. The results are displayed in the lowest of the three subwindows (Figure 4). Since the results of operations on histograms are not themselves histograms, in general, depicting them graphically as histograms requires an approximation to the true situation. However this is just a graphical convenience as the verified result may be graphically depicted in cumulative form (Figure 5). Any pane may be viewed in more detail as well (Figure 6).

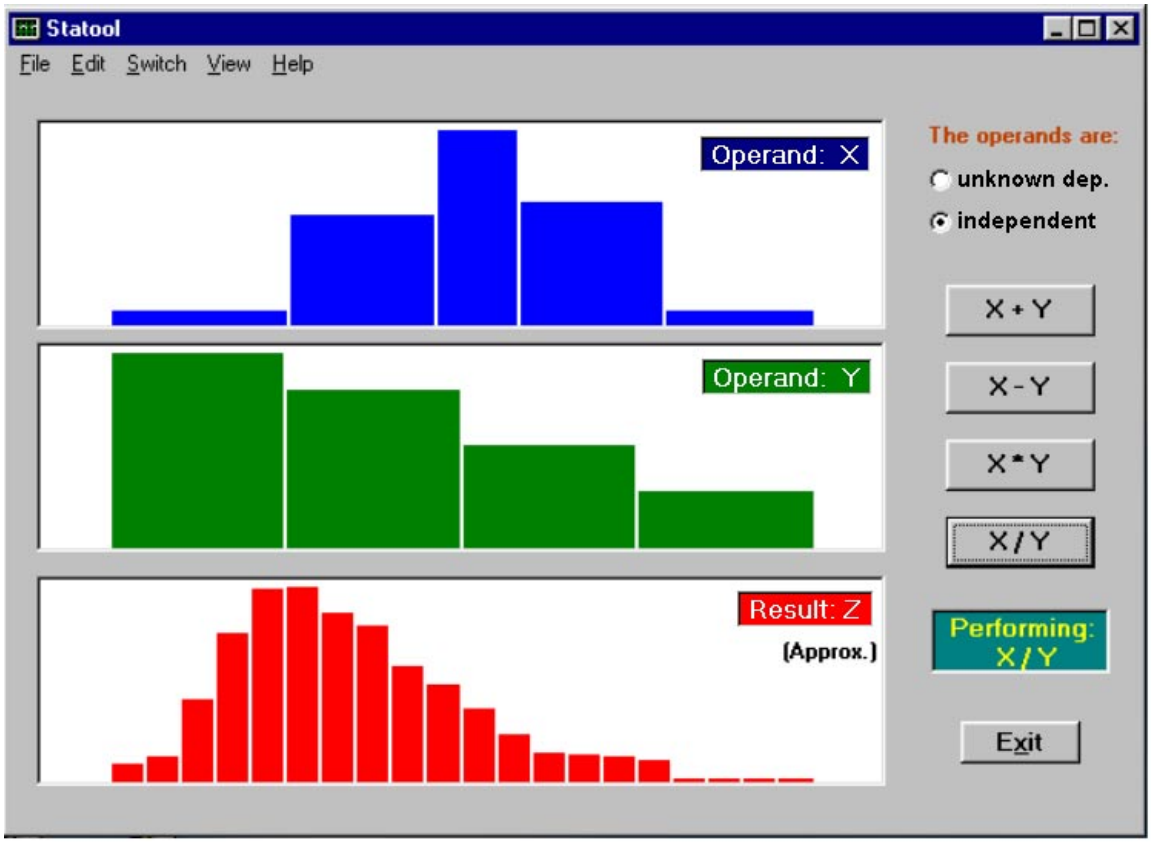


**Figure 4. Dividing operand X by operand Y produces a verified result Z, whose graphical depiction as a histogram is necessarily approximate. A correct graphical depiction would require either overlapping bars, or display of the cumulative form shown in Figure 5.**

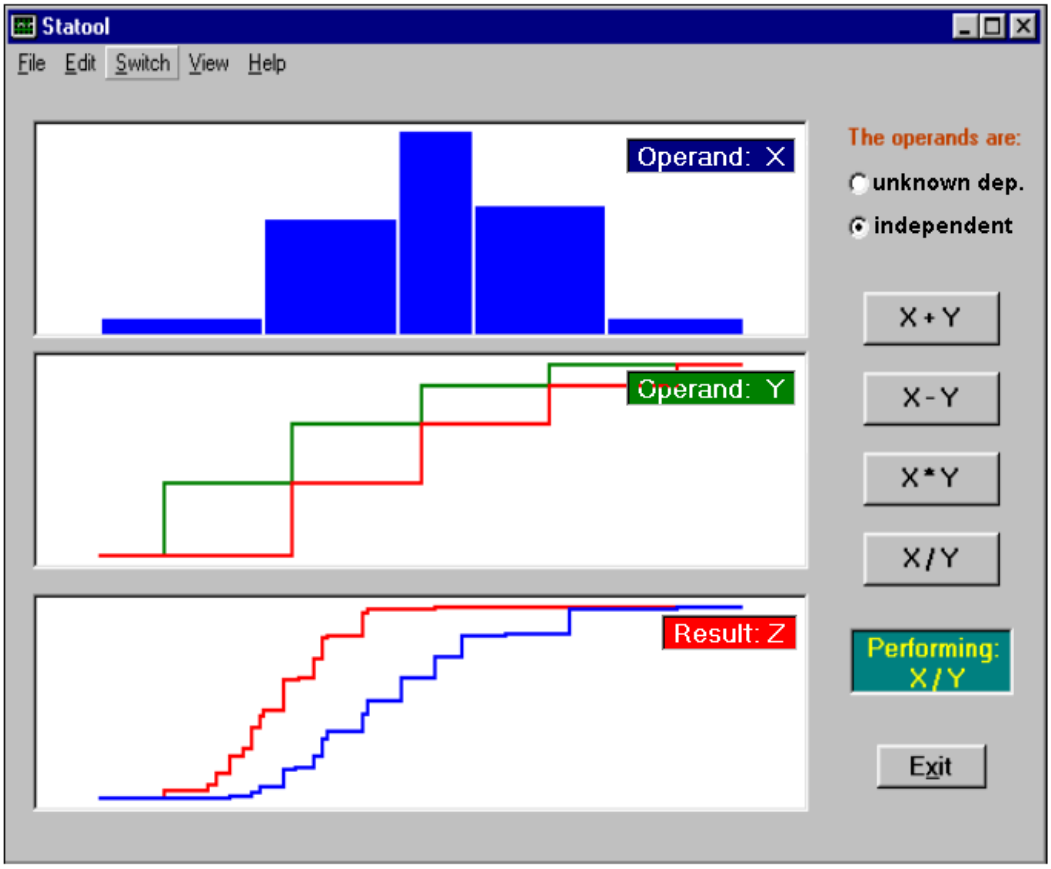


**Figure 5. The same data shown in histogram form in Figure 4 is shown here but with Y and Z in cumulative form. The cumulative form of a genuine histogram discretization of a PDF (Operand Y) is a pair of p-bounds that looks like two staircases in which the top bends of the lower curve touch the bottom bends of the upper curve. The cumulative form of the result does not in general obey that constraint, and hence cannot in general be displayed correctly as a histogram. It can be displayed correctly in cumulative form, as shown in the third subwindow. Numerical details of the cumulative form can also be shown (Figure 6).**

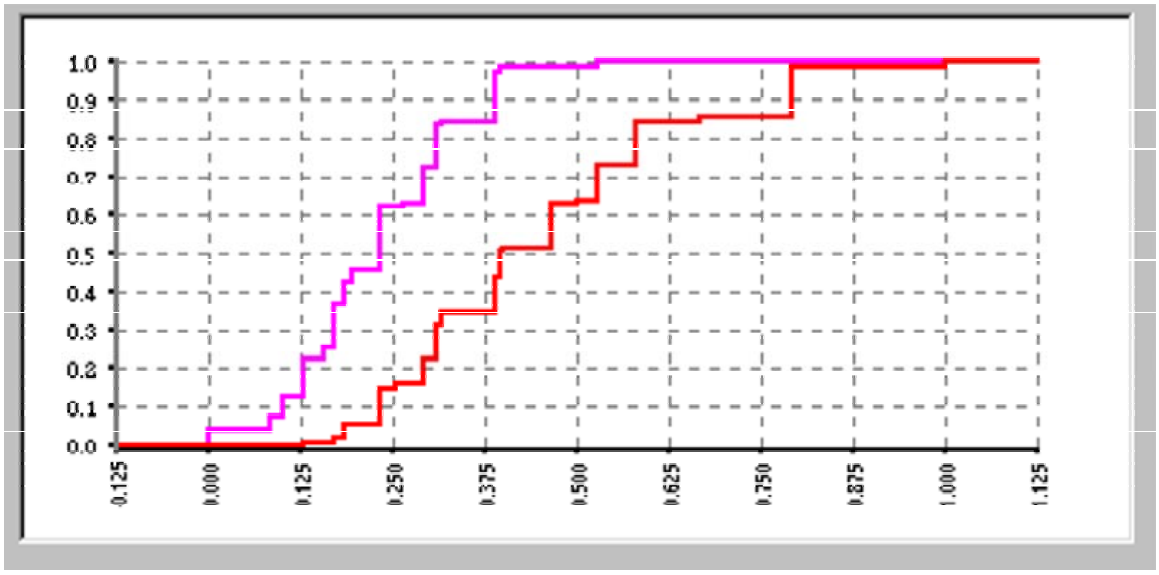


**Figure 6. P-bounds showing the bounds on a family of cumulative distributions resulting from an arithmetic operation on two discretized random variables. This is simply the detailed form of the Z subwindow from Figure 5.**

The use of p-bounds shows how results account for uncertainty in value caused by discretization and lack of knowledge about dependency relationships (Figure 7). Uncertainty is captured by showing upper and lower envelopes that bound the space of CDFs that could occur. Because the dependent case covers all types of dependency, the results it produces bound a superset of the space that is bounded by the analogous problem with the assumption of independence. In the independent case, the difference between the upper and lower bounding curves is due to discretization of the distributions of the operands (so that input histograms with more bars and consequently less coarse a discretization yield bounding curves that are closer together. In the case of unknown dependency, however, the difference is due to both discretization and the intrinsic uncertainty involved in refusing to assume any particular dependency relationship (Figure 8, which shows both independent and dependent curves for the otherwise same problem).

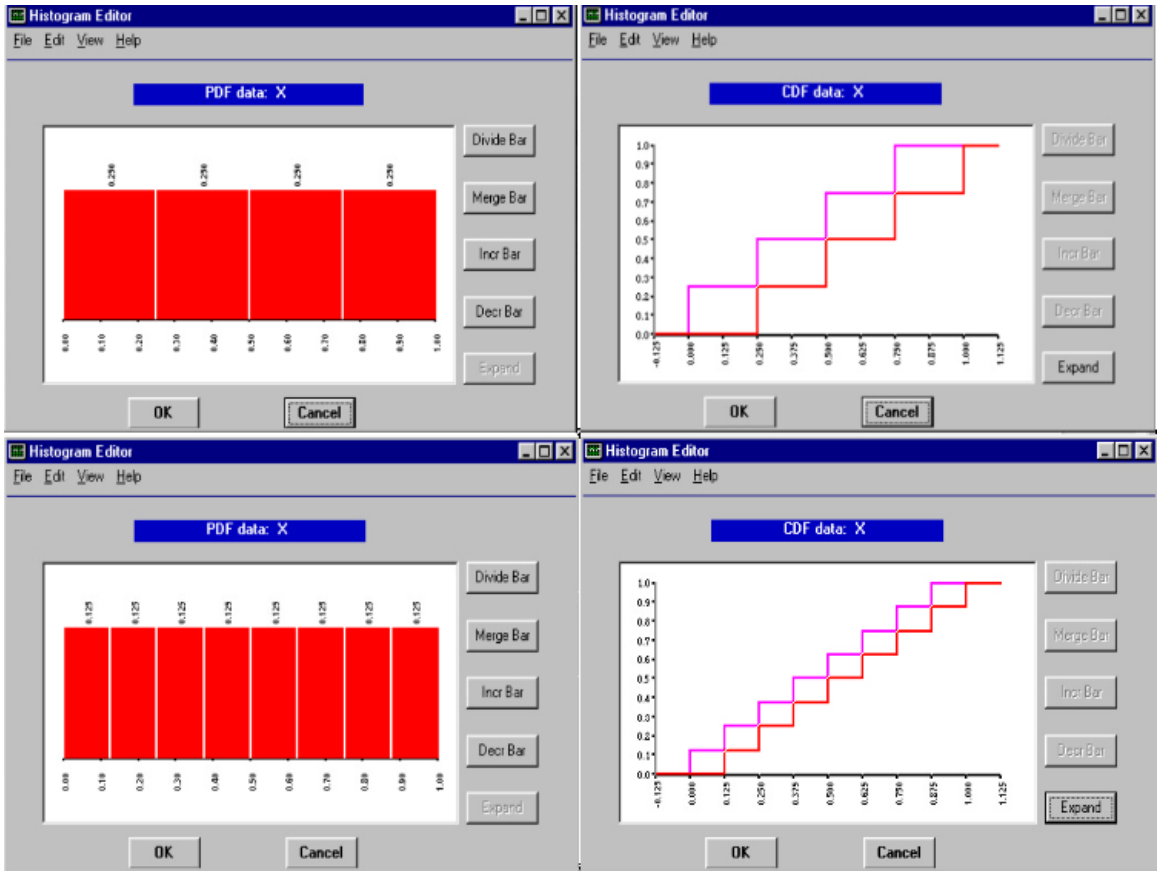


**Figure 7. Two histogram discretizations (left) and the p-bounds for their cumulative forms (right). The cumulative probability of a discretized distribution is fully defined at the point where two bars meet, so the corners of the upper and lower staircase curves touch at those points. The distribution of probability mass within the x-axis domain of any given bar is undetermined, and so could be concentrated on the left side of the bar, on the right side, or distributed over the domain of the bar in any less extreme way. If all masses are concentrated on the left of their respective bars the cumulative curve rises as fast as possible, giving the upper staircases; if concentrated on the right sides the cumulative curve rises as slowly as possible, giving the lower staircases. A finer-grained discretization (lower histogram and p-bounds) leads to narrower p-bounds than a coarser-grained discretization (upper histogram and p-bounds).**


**Figure 8. Overlay of two screen shots. The upper one shows the result of dividing the distribution in the top subwindow by the distribution in the middle subwindow, assuming they are independent. The lower one, partially obscured, shows the result under the weaker condition of unknown dependency. If assumed independent, narrower p-bounds result, while under no dependency assumption wider p-bounds result. Since the case of unknown dependence includes independence as one possibility, the p-bounds for the case of unknown dependence enclose the p-bounds for the case of independence.**

## Limitations

Statool has certain limitations which could be alleviated by extending the implementation. Our current research strategy is to identify application problems in order to develop the practical benefits of arithmetic on random variables of unknown dependency. The needs of these problems will then determine the priority of the extensions to be developed. Some possible extensions include the following.

1. Cascading operations. Currently the tool allows histograms as operands. When viewed as p-bounds, a histogram is reminiscent of two staircase-shaped curves in which the inward bends of the staircases touch (Figure 7). The result of an operation on them however in general produces staircases that do not touch, and are therefore not equivalent to any histogram. They can, however, be reformulated as an instance of a generalization of histograms: a set of intervals, each associated with some probability mass. This generalizes histograms by allowing intervals in the set to overlap, producing not

histogram but a "thicket." Thus, using the results of operations as inputs to other operations requires two things.

1. The histogram processing algorithms (convolution in the case of independence, linear programming in the case of unknown dependence) need to handle overlapping intervals. Fortunately, they can do this already since the convolution and LP algorithms make no assumption that the intervals in the joint distribution marginals are overlapping or not.
2. The result of operating on two histograms or other thickets under the unknown dependency condition is a pair of p-bounds, not a thicket. Therefore doing cascaded operations under unknown dependency will require a means of conversion from the staircase shaped p-bounds to thickets. This can be done, as exemplified by Figure 9.

2. Asymptotic pdf tails. The process of discretizing a pdf into a histogram does not presently allow for the case where a pdf tail trails off to plus or minus infinity. The solution is to allow the discretization to include open intervals with only one end point, the other end being infinite. This in turn would require the arithmetic operations to be defined on such intervals. Fortunately this is possible, e.g., $[1,\infty)+[1,2]=[2,\infty)$, $(-\infty,-1]*[-2,-1]=[1,\infty)$, $[1,2]/[-1,1]=(-\infty,\infty)$, etc.
3. Other arithmetic operations. Currently only +,-,*, and / are supported. Extending to other operations requires the operation in question to be defined on intervals. This has been reported for various other operations and so is not a difficult extension.
4. Expressions containing some combination of elementary operations. Currently evaluation of expressions can only be done as a cascade of operations. Thus (X+Y)-X would be done by calculating X+Y, then subtracting X from the result. This gives p-bounds that are too wide for the case of independent X and Y because of the excess width resulting from calculating (X+Y)-X on intervals X and Y. (For the case of unknown dependency between X and Y, cascading operations are not currently supported at all because conversion of staircase p-bounds into thickets is not currently supported.)

   Handling expressions in one step can be more convenient than cranking through a cascade of operations, and is essential in the case of expressions that lead to excess width if the excess width is to be recognized and removed. This requires an expression parser, the capability within the system to evaluate the expression on interval operands, and an expression evaluation algorithm that removes excess width when it occurs.

**Partial dependency**. In addition to the extensions just listed, which could be implemented with no further mathematical development, another direction of extension is of obvious interest but would require additional mathematical development. This is accounting for partial information about dependency. The system currently can calculate under the assumption of independence, or with no assumption about dependency (the case of unknown dependency). However partial information about dependency is often present in real problems, for example as correlation values. While partial information about dependency can sometimes be accounted for in ad hoc ways (an example is given in Berleant and Goodman-Strauss (1998), the general problem is non-trivial, yet occurs often in practical situations.

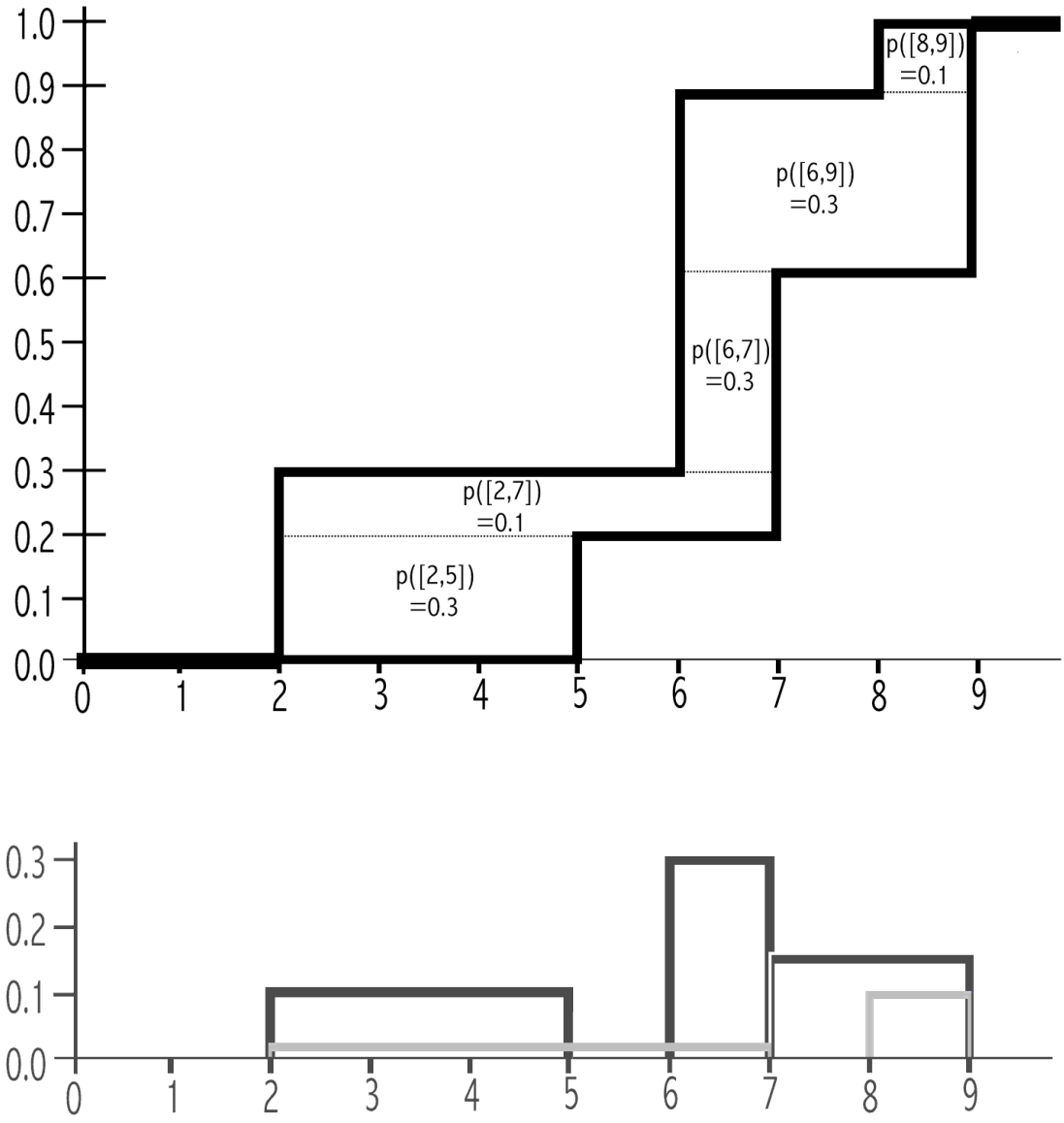


**Figure 9. Staircase shaped p-bounds (above) and the equivalent "thicket" of overlapping histogram bars (below). The distribution of mass within a given bar is undefined, with the flat tops being merely a graphical convenience. Any given distribution of mass within each bar leads to some curve enclosed by the p-bounds. To convert the p-bounds to a thicket of bars, the space between the p-bounds is partitioned into rectangles that span the space from left to right. Each rectangle represents an interval with bounds defined by its sides, and a corresponding probability mass defined by its height. Thus each rectangle defines a bar in a thicket of overlapping bars.**

## Availability

Statool may be downloaded from
http://class.ee.iastate.edu/berleant/home/Research/Pdfs/versions/statool/distribution/index.htm.
The source code is available in Visual Basic with DLLs written in C upon request.

Statool can be redistributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. The program is made available in the hope that it will be useful, but without any warranty; without even the implied warranty of merchantability or fitness for a particular purpose. See the GNU General Public License for more details.

## References


Berleant, D., Automatically Verified Reasoning with Both Intervals and Probability Density Functions, Interval Computations (1993 No. 2) 48-70.

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) 147-165.

Ferson, S., What Monte Carlo Methods Cannot Do, Human and Ecological Risk Assessment 2 (1996), pp. 990-1007.

Ferson, S. and M. Burgman. 1995. Correlations, dependency bounds and extinction risks. *Biological Conservation* 73:101-105.

Ferson, S., W.T. Root, and R. Kuhn, *RAMAS Risk Calc: Risk Assessment with Uncertain Numbers*. Applied Biomathematics, Setauket, New York, 1998. See also http://www.ramas.com/riskcalc.htm.

Frank, M.J., R.B. Nelson, and B. Schweizer, Best-Possible Bounds for the Distribution of a Sum – a Problem of Kolmogorov, Probability Theory and Related Fields 74 (199-211), 1987.

Ingram, G.E., E.L. Welker, and C.R. Herrmann, 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.

Kaplan, S., On the Method of Discrete Probability Distributions in Risk and Reliability Calculations – Applications to Seismic Risk Assessment, Risk Analysis, Vol. 1, No. 3, 1981, pp. 189-196.

Williamson, R. and T. Downs, Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds, International Journal of Approximate Reasoning, 4 (89-158) (1990).