1

AUTOMATICALLY VERIFIED ARITHMETIC ON PROBABILITY DISTRIBUTIONS AND INTERVALS

Daniel Berleant

Dept. of Computer Systems Engineering, University of Arkansas, Fayetteville, AR 72701, email: djb@engr.uark.edu

ABSTRACT

In this chapter we address two related problems:

1. Representing and operating on operands which are probability distribution functions; and
2. Representing and operating on operands when one is a distribution function and the other is an interval.

We discuss how to do these operations and get automatically verified results, using a method based on histograms. Histograms are composed of bars, each with an associated interval and probability mass. A histogram discretizes a probability distribution function in that each bar represents a portion of it, with the area of the bar equal to the area of the distribution function over that part of its domain corresponding to the interval-valued portion of the x-axis under the bar.

1 INTRODUCTION

The earliest use of histograms for representing and operating on distribution functions we have found appeared in 1968 (Ingram et al. [8]). That method was further developed by Colombo and Jaarsma in 1980 [4] and subsequently generated attention mostly in reliability analysis [9, 1, 5, 16] although the technique

*For completeness in presentation, this chapter contains some passages similar in content to passages appearing in [2].

itself is a general one. The method used was not automatically verifying. The method also served as a foundation for Kaplan's (1981 [10]) popular variation, which has generated over 50 citations in Science Citation Index over the years. However it is unclear how to further develop Kaplan's method into one which is automatically verifying. Moore (1984 [14]) independently developed another variation in which results are expressed as cumulative distribution functions (CDFs), which (because they are the integrals of probability density functions) begin at zero and rise to one over a typically S-shaped curve. R. E. Moore (1984 [14]) and A. S. Moore (1985 [13]) address non-trivial applications, but do not address automatic verification. A more recent variation on histogram arithmetic is described by Gerasimov et al. (1991 [6]).

The approximating nature of the method of Ingram et al. and Colombo and Jaarsma is due to assumptions about how the mass of any given bar of the histogram is distributed over the interval-valued portion of the x-axis under the bar. For example, because a bar is depicted graphically with a flat top one might assume that the associated mass is distributed evenly over the interval for that bar.

We have developed an automatically verified way of operating on histograms by modifying the original histogram based approach of Ingram et al. and Colombo and Jaarsma. In the next section, we describe how the discretization imposed by a histogram may itself be automatically verified in that the discretization weakens the statement made by the PDF, retaining correctness, rather than simply replacing the PDF with strong but approximate statement. Then, we describe how arithmetic operations on histograms can be automatically verifying. As a significant side benefit, automatically verified operations can be made on operands when one is a PDF and the other an interval, because both PDFs and intervals can be represented as histograms and arithmetic operations are done on histograms rather than on PDFs or intervals *per se.* While others have developed various methods for operating on PDFs such as using moments and defining special classes of PDFs, we are not aware of methods that work well when one of the operands happens to be an interval instead of a PDF.

2 CORRECTLY REPRESENTING PDFS AND INTERVALS WITH HISTOGRAMS

Histograms can be used to correctly represent both probability distribution functions (PDFs) and intervals. We take an interval I to be an incompletely specified statement about value. There is a probability of one that the value falls within the interval. Hence, an interval I defines a class of PDFs consisting of all PDFs $f(x)$ for which $f(x) = 0$ for $x \notin I$.

Let us examine first how to represent a PDF correctly using a histogram, and then how to represent an interval correctly using a histogram.

2.1 Correctly Representing a PDF with a Histogram

A histogram discretizes the underlying PDF. The discretization may be viewed as introducing information loss rather than error. To see this, consider a histogram discretization of a PDF to divide the PDF's domain on the x-axis into a set of intervals, each associated with a probability mass equal to the integral (area) of the PDF over that interval on the x-axis. The graphical depiction of a histogram is then simply a convention for describing this set of intervals and their associated masses, a set which we will call an *intermediate distribution.*

A histogram graphic (Figure 1, top) is user friendly but has the disadvantage that a histogram bar is conventionally shown with a flat top, which might give the impression that the probability mass associated with the interval for that bar is assumed to be distributed evenly over its range. In fact, in our technique no assumption is made about how the probability mass associated with an interval is distributed over that interval, hence a given histogram bar is consistent with any such distribution, not only the PDF that the histogram might have been intended to discretize but also an infinite number of other PDFs.

We can interpret a histogram as representing a PDF correctly, not approximately, by sacrificing some information about the PDF in representing it with a histogram. To do this, simply view the flat top of each bar as being an artifact. If one takes the perspective that a bar does *not* define how its area is distributed within its interval, then information loss by the discretization is implied since the original PDF *does* define exactly how its area is distributed. However, the limited statement the bar does make, that a certain area occurs under the PDF over a certain interval, is now correct. This is shown by Figure 2, which contains dissimilar PDFs which are all correctly describable using the same 2-bar histogram.

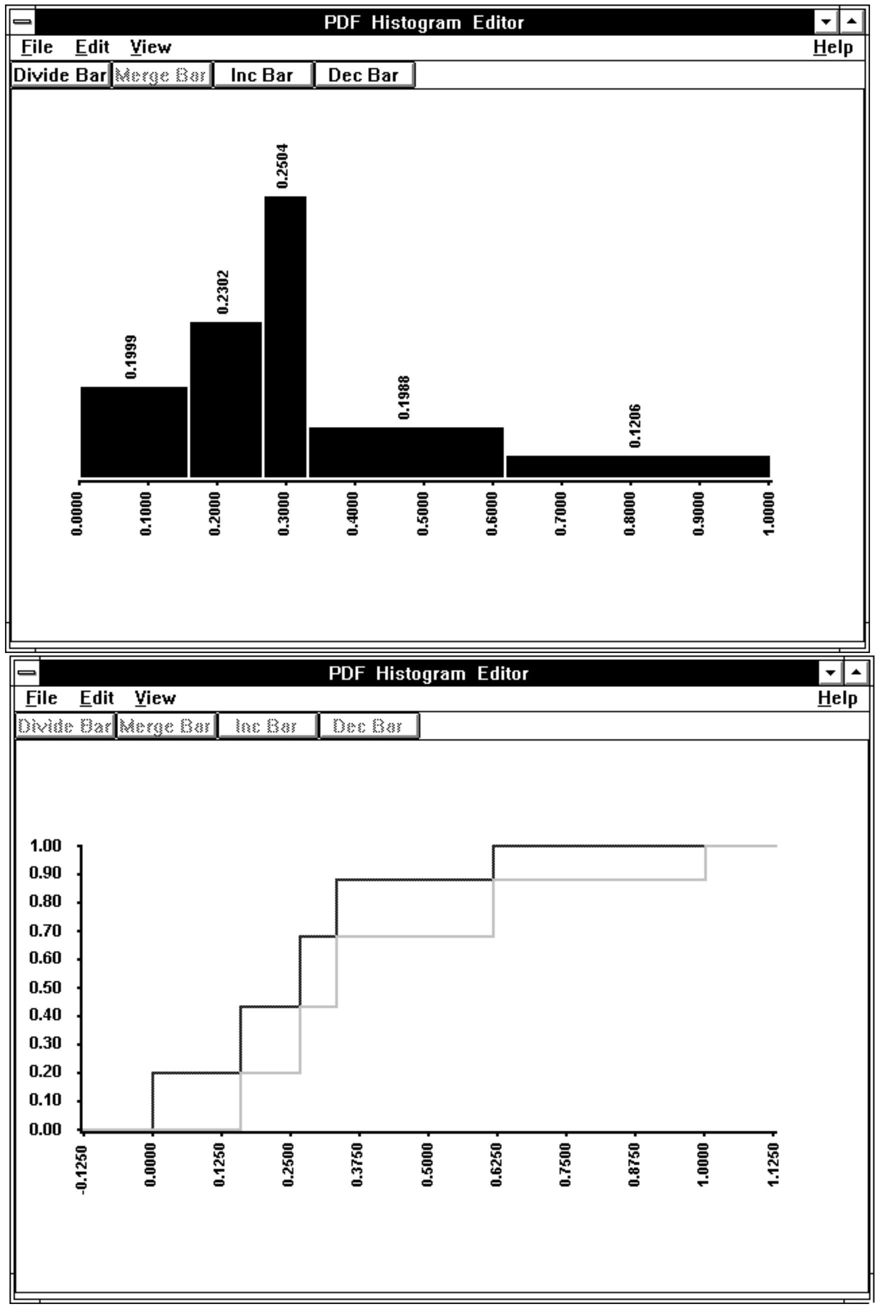


Figure 1 A histogram and its two corresponding CDF bounds (one solid and one dotted).

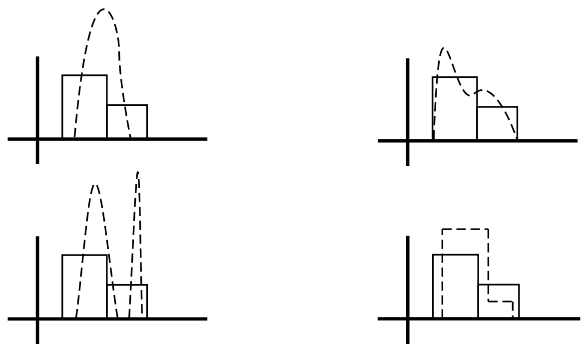


Figure 2 Four out of an infinite set of probability density functions representable with the same histogram. A histogram bar spans an interval I, and a height. Consequently it has an area which encodes a probability mass. This area equals the area (and the probability mass) of the PDF over interval I. (Figure from [2].)

Integrations of Histograms can Also be Shown Graphically.

A histogram (representing a family of PDFs) can be integrated to produce two cumulative distribution functions (CDFs) that bound the family of CDFs corresponding to the family of PDFs. The higher of the two assumes the extreme case where the probability mass associated with each bar of the histogram is distributed in a very skewed way — all at the lower bound of its associated interval. The lower of the two bounding CDFs assumes the other extreme, that each probability mass is concentrated at the upper bound of its corresponding interval. Any other of the infinite number of distributions consistent with the histogram, when integrated, results in a curve that stays within the two stair-step shaped bounding CDFs exemplified in Figure 1. The space between these bounding CDFs reflects the information loss associated with the discretization.

2.2 Correctly Representing an Interval with a Histogram

We take an interval to bound a value such that it has probability one of being within the interval, and probability zero of being outside it. An interval is represented in our technique as a histogram with one bar. To represent the interval using bounding CDFs, we must determine the space of CDFs consistent with the interval. The two most extreme CDFs that bound this space are the one that rises faster than any other CDF consistent with the interval, and the one that rises slower than any other CDF consistent with the interval. The one that rises fastest will correspond to the PDF which is an impulse of mass 1 at the interval's lower bound (that is, the variable's value is certainly equal to the interval's lower bound). Likewise, the one that rises slowest will correspond to the PDF which is an impulse of mass 1 at the interval's upper bound. The interval is therefore representable as a family of CDFs whose two bounding CDFs each have one "step." Figure 3 shows an interval and its representation as a bounded family of CDFs; the CDF that bounds this family from above jumps from zero to one at the interval's lower bound, and CDF that bounds it from below jumps from zero to one at the interval's high bound.

3 ARITHMETIC OPERATIONS

We introduce arithmetic on PDFs in the context of an example (Figure 4a–d) (a more abstract presentation of the method appears in [2]). The interval arithmetic operation, whichever one it is, is applied to the interval for a bar i of operand 1 (Figure 4a) and the interval for a bar j of operand 2 (Figure 4b), producing a result interval I_{ij}, for every possible i and j (Figure 4c). In other words, the operation is applied to each member of the Cartesian product of the bars in operand 1 and the bars in operand 2. The result is a set of intervals I_{ij} collectively describing the result, called an *intermediate distribution* because it is neither a PDF nor a CDF, but is instead a list of intervals and their associated probability masses. The probability mass associated with each I_{ij} in this result is the probability that the value of operand 1 is in bar i, which we will call p_i, and that the value of operand 2 is in bar j, which we will call p_j. Assuming independent operands, the probability associated with resulting interval I_{ij} is then p_ip_j, as shown in the table of Figure 4c. The independence assumption is a limitation which could be eliminated by extending the method to non-independent operands whose joint distribution is known.

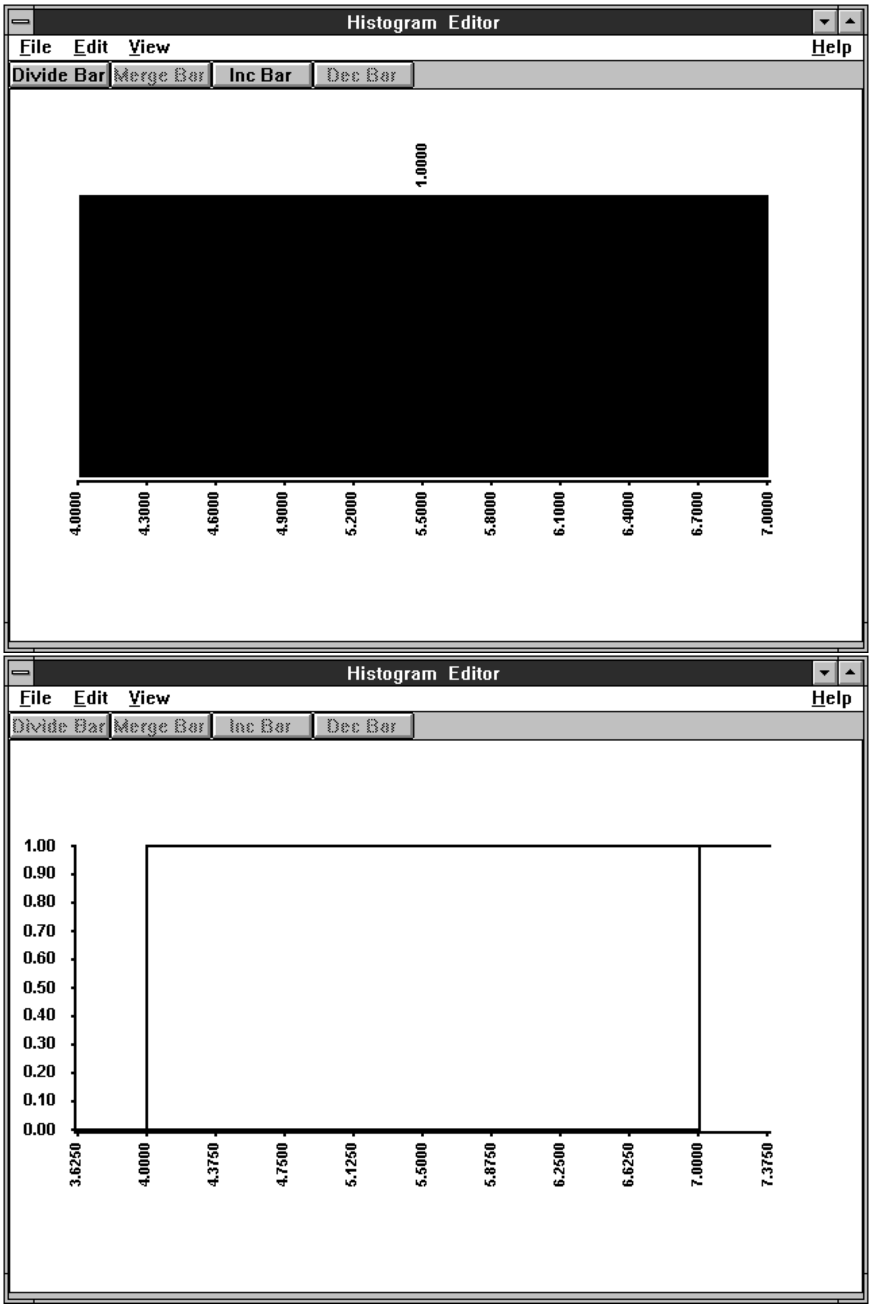
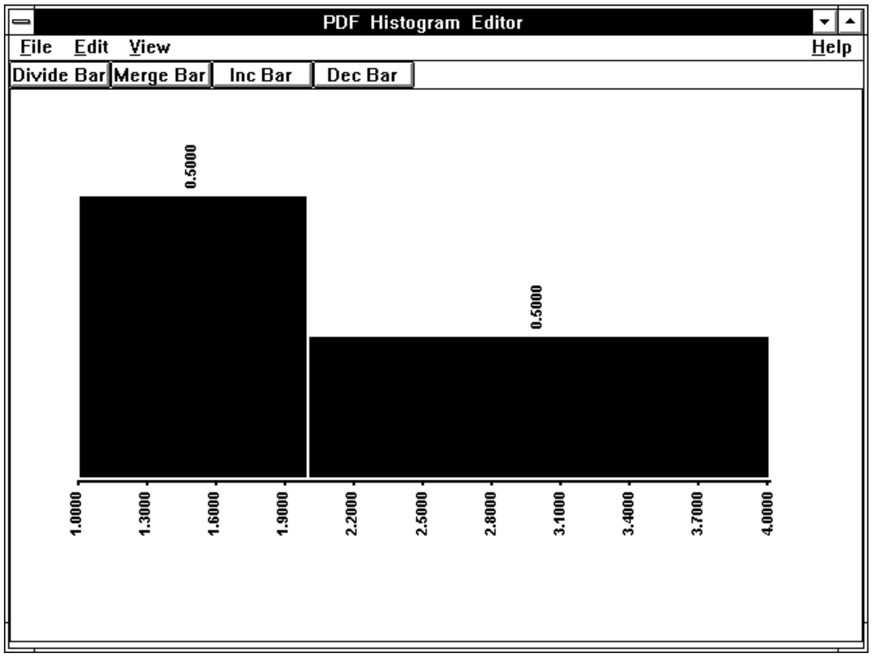
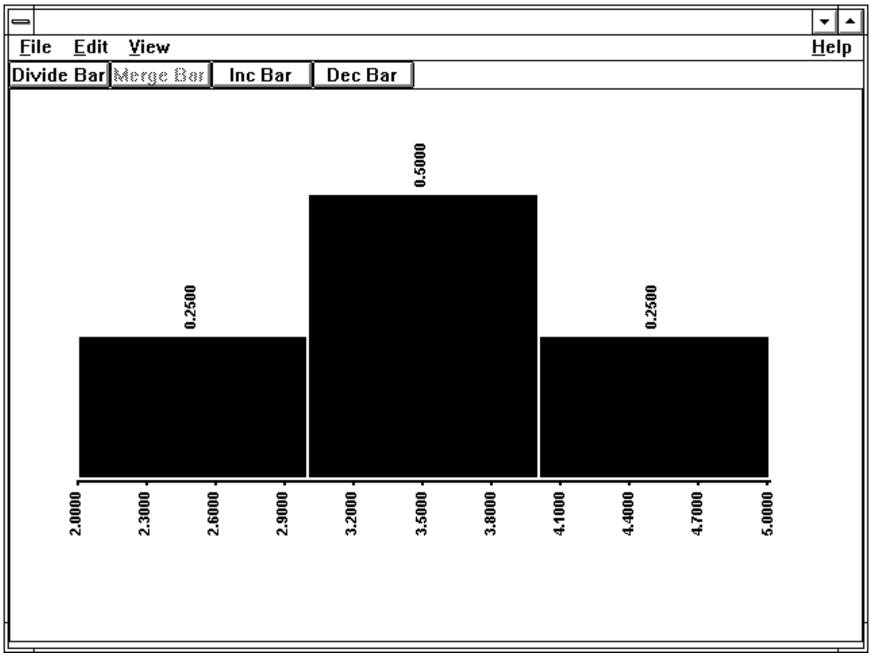


Figure 3 The interval $[4, 7]$ and its representation as a 1-bar histogram (top) and as two CDFs bounding the family of all CDFs consistent with the interval (bottom).



4a: Let this histogram describe uncertain value X. It depicts the intermediate distribution $\{p([1,2]) = \frac{1}{2}, p([2,4]) = \frac{1}{2}\}$.

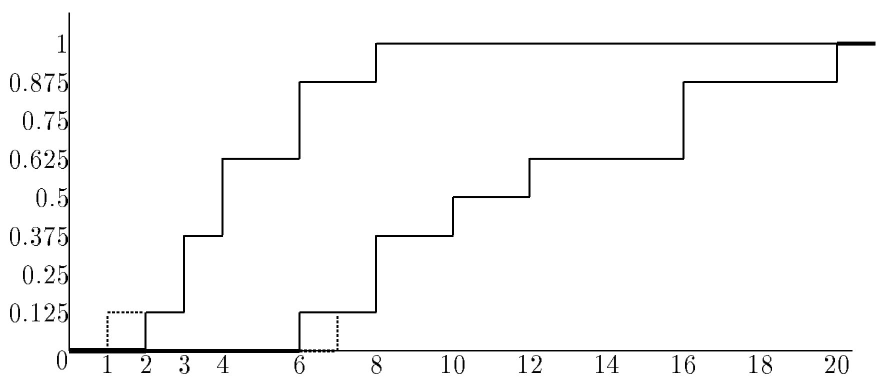
Figures 4a–d illustrate the multiplication of two operands. The user created two histograms X (Figure 4a) and Y (Figure 4b). These histograms are represented internally as intermediate distributions, that is, lists of intervals and associated probability masses. The intermediate distributions are used to calculate $X \times Y$, whose intermediate distribution is shown in the last column of Figure 4c. Integrating this result produces the CDF bounds shown in Figure 4d. (The dotted lines in the lower left portion of that figure exemplify the effect of excess width, see section 3.2.)



4b: Let this histogram describe uncertain value Y. It depicts the intermediate distribution $\{p([2,3]) = \frac{1}{4}, p([3,4]) = \frac{1}{2}, p([4,5]) = \frac{1}{4}\}$.

Cartesian product term #		Operand$_1$ (X_i)	Operand$_2$ (Y_j)	Result intermediate distribution
#1	interval probability	$[1,2]$ $1/2$	$[2,3]$ $1/4$	$[2,6]$ $1/8$
#2	interval probability	$[1,2]$ $1/2$	$[3,4]$ $1/2$	$[3,8]$ $1/4$
#3	interval probability	$[1,2]$ $1/2$	$[4,5]$ $1/4$	$[4,10]$ $1/8$
#4	interval probability	$[2,4]$ $1/2$	$[2,3]$ $1/4$	$[4,12]$ $1/8$
#5	interval probability	$[2,4]$ $1/2$	$[3,4]$ $1/2$	$[6,16]$ $1/4$
#6	interval probability	$[2,4]$ $1/2$	$[4,5]$ $1/4$	$[8,20]$ $1/8$

4c: Multiplying the intermediate distributions that describe the intermediate distributions for X and Y leads to another intermediate distribution, shown in the last column. (This intermediate distribution contains overlapping intervals, as is typical of intermediate distributions resulting from arithmetic operations.)



4d: Integrating the resulting intermediate distribution in the last column of Figure 4c produces these stair-step shaped, bounding CDFs. (Figure from [2].)

Since no assumption may be made about the distribution of mass within any interval in either operand, no assumption may be made about the distribution of mass within any interval of the result. Therefore the integral of the intermediate result collection cannot be fully defined. However we can bound that integral with upper and lower CDFs. These CDFs bound a space of CDFs containing all CDFs corresponding to some distribution of the intermediate resulting interval probability masses over their respective intermediate resulting intervals.

This is how the bounding CDFs are derived. For each member of the resulting intermediate distribution, its integral can rise faster or slower, depending on its distribution of probability mass. By assuming each member's integral rises as fast as possible we get one bounding CDF. By assuming each member's integral rises as slow as possible we get the other bounding CDF.

Illustrating the integration of the resulting intermediate distribution of Figure 4c to produce Figure 4d, first consider row #1. The part of the resulting intermediate distribution derived in that row specifies probability $\frac{1}{8}$ distributed somehow over interval $[2,6]$. If that $\frac{1}{8}$ mass was distributed by having it all over the lower bound of the interval, 2, then the integral of the intermediate distribution comprising the result would rise immediately to $\frac{1}{8}$ as the integral curve passed the value 2 on the x-axis. This constitutes the fastest possible rise that interval $[2,6]$ could lead to. Looking next at the $\frac{1}{4}$ probability for row #2, the fastest possible rise it could cause would be if that probability mass was concentrated over the lower bound of its interval, 3, and so the CDF would discontinuously rise above its previous level by $\frac{1}{4}$ at a value of 3 on the x-axis. If all the other rows in Figure 4c are treated similarly, the integral will rise at the maximum possible rate and form the upper bounding CDF depicted in Figure 4d.

The lower bounding CDF is obtained the same way, except that the probability masses associated with the rows of the table are assumed to be concentrated at the upper bounds of their associated intervals, instead of the lower bounds as before. The resulting bounding CDFs are automatically verified, because any CDF corresponding to a PDF consistent with the histogram will fall within those bounds.

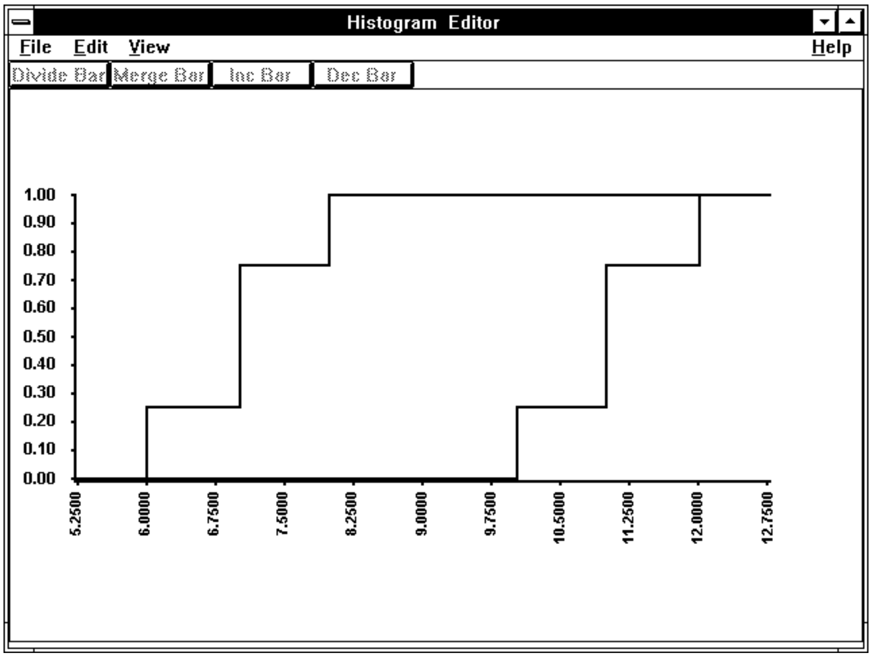
An approach to using CDF bounds in computer systems administration appears in Post and Diltz (1986 [15]) under the rubric of the *stochastic dominance* field [12], which recognizes the usefulness of bounded families of CDFs in fields such as finance. Another area of related work is robust statistics.

3.1 Combining a PDF with an Interval

Since an interval can be represented as a histogram with one bar, exactly the same process as is used to combine two PDF operands can be used to combine a PDF operand and an interval operand. As an example, consider describing Z, where $Z = X + Y$, X is any PDF discretizable as the histogram shown earlier in Figure 4b, and Y is the interval $[4, 7]$ (shown earlier as a histogram in Figure 3, top). Figure 5a shows the Cartesian product table with the intermediate distribution of the result shown in the last column. Integrated, this may be correctly depicted graphically, as shown in Figure 5b.

Cartesian product term #		Operand$_1$ (X_i)	Operand$_2$ (Y_j)	Result intermediate distribution
#1	interval	$[2, 3]$	$[4, 7]$	$[6, 10]$
	probability	1/4	1	1/4
#2	interval	$[3, 4]$	$[4, 7]$	$[7, 11]$
	probability	1/2	1	1/2
#3	interval	$[4, 5]$	$[4, 7]$	$[8, 12]$
	probability	1/4	1	1/4

5a: Adding X and Y where X is a histogram representation of a PDF, and Y is a histogram representation of an interval.



5b: CDF bounds on the result of adding a PDF and an interval.

Figure 5. Adding a PDF and an interval. Since the histogram representation can express both PDFs and intervals, the method described can do arithmetic on operands regardless of whether an operand is a PDF or an interval.

3.2 Excess Width

Interval calcuations are an important component of the method, and excess width is a possibility under some circumstances. For example, the desired operation on two histograms may be an expression that, when evaluated on the interval for a bar in one histogram and the interval for a bar on the other, results in an interval containing excess width. The resulting set of intervals would then contain intervals with excess width.

When intervals in the result of an operation have excess width, the intervals that constrain how probability masses are distributed are too wide, so that the process of finding the two most extreme distributions of mass within an interval to help form the upper and lower bounding CDFs produces extreme distributions that are too extreme. For example, concentrating the mass at an interval lower bound that is too low, when integrated to form a bounding CDF, makes the CDF rise too soon. Likewise, an upper bound that is too high leads to a bounding CDF that rises too slowly. As an example, consider row #1 of Figure 4c. The interval is $[2, 6]$, but suppose the calculation that produced it added excess width and instead estimated the interval as $[1, 7]$, wider than it really is. Then one extreme distribution of its mass of $\frac{1}{8}$ would have it concentrated at 1, and the other extreme distribution would have it concentrated at 7. This would widen the bounding CDFs as shown by dotted lines in Figure 4d. In general, excess width in calculations of intervals in the result of an operation creates weaker bounding CDFs. This means results are weaker, but still automatically verified.

3.3 Other Related Methods

One approach to operating on PDFs is to address only a limited class of PDFs, such as the class of lognormal curves. This method has the disadvantage over the method described here in that it does not apply to arbitrary PDFs.

A second approach is to use moments of PDFs to calculate moments of the result of combining them. This has usually meant approximating, although Rushdi and Kafrawy [16] present a variation providing "readily quantifiable accuracy that can be arbitrarily improved." However they do not address the issue of operations where one operand is a PDF and the other an interval.

A recent constraint satisfaction approach to the problem of operating on PDFs is given by Hyvönen (1995 [7]), who use a histogram-like representation to express envelopes around the PDF, intead of around the CDF as in the present work.

The most widely known and used method is the Monte Carlo method. A comparison with Monte Carlo is next (Table 1 summarizes).

Disadvantages of the automatically verified histogram method compared to the Monte Carlo method

The automatically verified histogram method does not presently handle dependent operands, whereas the Monte Carlo approach can, if the dependency is known. However, the automatically verified histogram method could be ex-

	Automatically verified histogram method	Monte Carlo methods
Handles dependent inputs?	(not currently)	$\surd$
Automatically verifying?	$\surd$ (excess width possible)	$\times$ (insufficient width almost certain)
Handles interval inputs?	$\surd$	$\surd$
Handles PDF inputs?	$\surd$	$\surd$
Handles PDFs and intervals?	$\surd$	$\times$

Table 1 A comparison of the Monte Carlo and the automatically verified histogram methods.

tended to handle dependent inputs in that a joint distribution could be used to get the probabilities of the intervals in the intermediate distriabution resulting from an operation, instead of simply assuming independence and getting the probabilities of the intervals in the result by multiplying the probabilities of the operand intervals as in Figure 4c.

Advantages of the automatically verified histogram method compared to the Monte Carlo method

Monte Carlo methods work by randomly choosing input values. Many sets of input values are chosen this way and used to get many corresponding output values. For inputs known to fall within intervals, the range of outputs generated by a sufficiently large random sample of possible inputs is used to estimate the range of possible outputs. However there is a possibility that those inputs that truly result in the most extreme outputs were never tried, so that the observed range of outputs is a subset of the actual range of outputs possible. Thus the observed range of outputs is likely to be too narrow, and hence the results of the Monte Carlo process are unguaranteed.

For inputs that are known sufficiently so that they can be characterized by PDFs, the inputs would be randomly sampled according to the PDFs so that for sufficiently large numbers of samples the distribution of samples will approximate the PDF used to generate them. The observed distribution of outputs corresponding to such a set of inputs then approximates the true PDF for the output. The likely deviation from that true PDF can be statistically characterized as an envelope around the CDF of the observed PDF. That envelope has a confidence level associated with it (e.g. Kolmogorov 1941) [11], not a guarantee. However, the automatically verified histogram method does provide a guarantee for the enveloping CDF bounds it produces.

Another advantage of the automatically verified histogram method applies when one input variable is a PDF and the other an interval. Monte Carlo methods have difficulty handling such a problem because of the difficulty of sampling an input space consisting of both intervals and PDFs. However the automatically verified histogram method works just as well when one operand is a PDF and the other an interval as it does when both operands are PDFs.

Acknowledgements

The author thanks Hang Cheng for implementing the method in a software tool [3], and Bill Walster for pointing out that the method could be extended to dependent operands if their joint distribution is known.

REFERENCES

[1] S. Ahmed, R. E. Clark, and D. R. Metcalf, "A Method for Propagating Uncertainty in Probabilistic Risk Assessment," *Nuclear Technology*, 1982, Vol. 59, pp. 238–245.

[2] D. Berleant, "Automatically Verified Reasoning with Both Intervals and Probability Density Functions," *Interval Computations*, 1993, No. 2, pp. 48–70.

[3] H. Cheng and D. Berleant, "A Software Tool for Automatically Verified Operations on Intervals and Probability Distributions," submitted to the International Journal *Reliable Computing*.

[4] A. G. Colombo and R. J. Jaarsma, "A Powerful Numerical Method to Combine Random Variables", *IEEE Transactions on Reliability*, 1980, Vol. R-29, No. 2, pp. 126–129.

[5] F. Corsi, "Mathematical Models for Marginal Reliability Analysis," *Microelectronics and Reliability*, 1983, Vol. 23, No. 6, pp. 1087–1102.

[6] V. A. Gerasimov, B. S. Dobronets, and M. Yu. Shustrov, "Numerical Calculations for Histogram Operations and their Applications," *Avtomatika i Telemekhanika*, 1991, Vol. 52, No. 2.

[7] E. Hyvönen, *Constraint Reasoning on Probability Distributions,* manuscript (1995), author is at VTT Information Technology, P.O. Box 1201, 02044 VTT, Finland.

[8] G. E. Ingram, 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.

[9] P. S. Jackson, R. W. Hockenbury, and M. L. Yeater, "Uncertainty Analysis of System Reliability and Availability Assessment", *Nuclear Engineering and Design*, 1981, Vol. 68, 1981, pp. 5–29.

[10] S. Kaplan, "On the Method of Discrete Probability Distributions in Risk and Reliability Calculations — Application to Seismic Risk Assessment," *Risk Analysis*, 1981, Vol. 1, No. 3, pp. 189–196.

[11] A. Kolmogoroff (*a.k.a.* Kolmogorov), "Confidence Limits for an Unknown Distribution Function," *Annals of Mathematical Statistics*, 1941, Vol. 12, No. 4, pp. 461–463.

[12] H. Levy, "Stochastic Dominance and Expected Utility: Survey and Analysis," *Management Science*, 1992, Vol. 38, No. 4, pp. 555–593.

[13] A. S. Moore, "Interval Risk Analysis of Real Estate Investment: A Non-Monte Carlo Approach," *Freiburger Intervall-Berichte*, 1985, Vol. 85/3, pp. 23–49.

[14] R. E. Moore, "Risk Analysis Without Monte Carlo Methods," *Freiburger Intervall-Berichte*, 1984, Vol. 84/1, pp. 1–48.

[15] G. V. Post and J. D. Diltz, "A Stochastic Dominance Approach to Risk Analysis of Computer Systems," *Management Information Systems Quarterly*, 1986, Vol. 10, pp. 363–375.

[16] A. M. Rushdi and K. F. Kafrawy, "Uncertainty Propagation in Fault-Tree Analyses Using an Exact Method of Moments," *Microelectronics and Reliability*, 1988, Vol. 28, No. 6, pp. 945–965.