Preprint draft of: D. Berleant, G. Anderson, and C. Goodman-Strauss, "Arithmetic on bounded families of distributions: a DEnv algorithm tutorial," chapter 9 in C. Hu, R.B. Kearfott, A. de Korvin, and V. Kreinovich, eds., *Knowledge Processing with Interval and Soft Computing*, pp. 183-210, Springer, 2008 (ISBN 978-1-84800-325-5).

1

Arithmetic on Bounded Families of Distributions: a DEnv Algorithm Tutorial

Daniel Berleant and Gary Anderson

University of Arkansas at Little Rock, 2801 S. University Ave., AR 72212, USA
{jdberleant,gtanderson}@ualr.edu

Monte Carlo analysis is traditionally used in risk analysis to model uncertainty in the values of inputs of various kinds, such as initial conditions and variables. While Monte Carlo has proven useful, extensive experience has revealed limitations in the technique. These limitations have motivated new techniques that overcome those limitations. This chapter focuses on an alternative approach, the DEnv algorithm. We begin by briefly discussing limitations of Monte Carlo simulation, followed by ways of attempting to address these limitations within the Monte Carlo paradigm. Then we discuss the DEnv (from Distribution Envelopes) algorithm, a technique for working with bounded families of probability distributions.

1.1 Motivation: Monte Carlo simulation and its limits

It is useful to start with a critical look at Monte Carlo simulation, because the benefits of bounded families of distributions can best be appreciated in the context of the limitations of the traditional Monte Carlo method. There are two limitations that are especially significant in motivating use of bounded families of distributions in certain problems. These are described in the next two subsections.

1.1.1 When knowledge is insufficient to specify a probability distribution for a model variable

If some variable is uncertain, that uncertainty can often be modeled with a distribution function. However if insufficient information exists to specify the exact shape of that distribution function it is impossible to draw the samples needed by Monte Carlo simulation, unless either some distribution is arbitrarily assumed to apply, or the variable is described by an interval instead of a distribution function. An arbitrary distribution (e.g. a normal or "bell" curve) might be assigned to a variable so that samples could be drawn from it, but while this would enable Monte Carlo analysis to proceed the cost would be in making an unjustified assumption about the variable. Unjustified assumptions about model variables tend to imply results that, literally, are also unjustified.

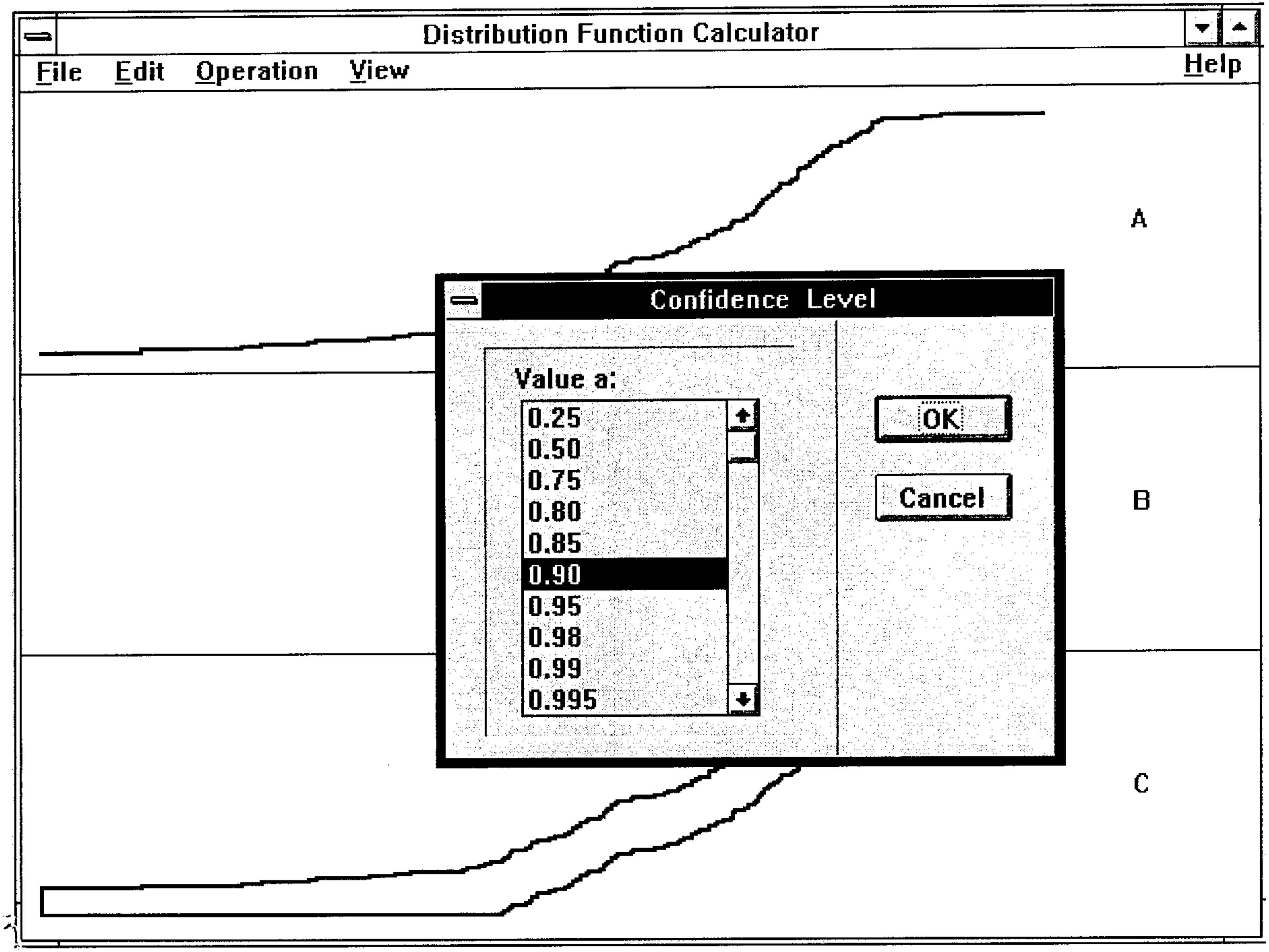
To get justifiable results, given a variable with an unknown distribution, one might choose to bound it by using an interval to describe its range (possibly excluding the tails of the distribution if such a move is reasonable given the problem). Then this interval could be sampled, leading to results that bound the range of values of the outputs. However an interval is a relatively weak characterization of a variable that ignores information that may be available, such as variance and mean, that could potentially be used to help characterize the outputs.

Let us consider two examples. In the first, the information available is insufficient to specify a single distribution. In the second, and interval is suitable for describing uncertainty about the available information.

> *Example 1:* Kolmogorov (1941) showed that a distribution obtained from a limited number of data points is likely to be significantly wrong, and that confidence limits (in the form of bounds around the nominal distribution defined by the data) are more appropriate. Figure 1.1 shows an example of a distribution function and its confidence limits. Frame A, at top, shows a cumulative distribution obtained from a set of data points. The s-curve shown rises unevenly due (one might reasonably speculate) to random noise in the data, although in principle the unevenness might actually accurately reflect the underlying random variable. Frame C, at bottom, shows probability bounds that describe the confidence limits of the curve of frame A at the 0.9 probability level. The true distribution (which could be obtained from a limitless number of sample data) will fall within the bounds of the confidence limits with a probability of 0.9. Put another way, there is a 0.1 probability that the true distribution will cross outside the enveloping bounds shown at least once.

Fig. 1.1. A distribution derived from data, top, and its confidence limits, bottom (from Berleant and Cheng, 1998, Fig. 6.).

> *Example 2:* A manufacturer of thermostats or some other measurement or control device might state limits on the measurement error and the controlled quantity. In this instance, information exists about the range of a variable but not about its distribution within that range. Intervals would be appropriate here for expressing uncertainty



Value a:

0.25
0.50
0.75
0.80
0.85
0.90
0.95
0.98
0.99
0.995

Fig.1d

because they state lower and upper bounds. Continuing the thermostat case, if the temperature setting s is 67°, the manufacturer states that settings are accurate to ±1°, and the manufacturer also specifies a hysteresis of ±1° (that is, the heater turns on at $s - 1 = 66°$ and turns off at $s + 1 = 68°$), then the actual temperature can be inferred to be within $(67 \pm 1 \pm 1)°$, or $[65, 69]°$. However its distribution within that range cannot be determined.

1.1.2 Lack of full knowledge about the dependency relationships among variables

Suppose two variables have no significant relation to each other. For example, the price of oranges has no significant relation to the number of sunny days per year in Seattle, USA. If distribution functions are available for both variables, each may be sampled to provide pairs of numbers to use in a simulation model, without fear that the value of a sample of one variable affects the distribution function that should be used to generate samples of the second variable. This is convenient both in implementability and in ease of modeling.

Another convenient situation is if the value of one variable completely determines the value of a second variable. For example, the number of sunny days per year completely determines the number of non-sunny days per year, and both values might be used in a simulation model of, say, utilization of tourist attractions (of which some would be more attractive on sunny days and others on non-sunny days). In this situation a sample drawn from one variable determines the sample to use of the other, and a simulation requiring both variables is both relatively simple and implementable.

However, a third situation often occurs which presents a problem. For example, an agricultural model of production might incorporate as variables both the price of the product and the number of sunny days in growing areas. It is likely that the values of those variables will be related in some way (i.e. not independent), but that this dependency is less than total (i.e. the value of one does not fully determine the other). Unless the joint probability distribution is known, which amounts to knowing exactly what distribution to sample for one variable given what value was sampled for the other, it is not possible to properly generate a sample value of one variable given a sample value of the other.

1.1.3 Overcoming the limitations of Monte Carlo while staying within the paradigm

Let us look at how the Monte Carlo approach may be made usable in situations in which the just-mentioned two limitations occur. Later this will help illustrate the advantages of a better approach, bounded families of distributions. The next two subsections address the two limitations.

Knowledge insufficient to specify the probability distribution of a model variable

Often a generic "reasonable" distribution will be used to model such a variable, for example a normal distribution. This permits a Monte Carlo model to be fully specified and therefore a simulation to be run. However such an unjustified assumption about a model variable of course decreases the dependability of conclusions drawn from simulating the model.

Because potentially untrue assumptions can lead to problematic conclusions, we might wish to express only actual facts about variables. For example, we might model variables as intervals (i.e. ranges extending from minimum to maximum plausible values). If we expressed all variables this way, then a Monte Carlo simulation could be performed based on picking sample points randomly within those intervals for each simulation run. The results from many simulation runs would then be combined to give intervals describing ranges for the outputs. Unfortunately it would say nothing about the shapes of their distributions, merely giving estimates of the ranges of their supports.

What if only some variables needed to be described using intervals because distributions were available for the others? Simulating Monte Carlo models which mix some variables that are interval-valued and others that are distribution function-valued is less straightforward than if all were intervals or all were distributions. It would be easier to substitute, for each distribution in such a mixed model, an interval bounding the range of values permitted by the distribution. A disadvantage of this approach is that using intervals for variables for which distribution functions are known means ignoring available information. While the conclusions drawn may be sufficient in some situations, they will tend to be weaker than if the distribution information available was used instead of ignored.

> *Example 3:* We model whether a robotic vehicle can pull a cart containing cargo up a slope without its wheels slipping against the slope surface (Fig. 1.2), rendering it unable to complete its task. This example can be applied to specific situations such as a robot moving cargo from an airplane drop to a central location, cargo transportation in rough and/or dangerous terrain, autonomous construction of bridges or other structures, and so on.

Fig. 1.2. Pioneer AT robot pulling a loaded cart up a hill.

> The frictional force between the surface and the drive wheels of the robot must exceed the gravitational force pulling the cart down the incline (see Fig. 1.2). The force of gravity on the cart is $m_{cart} * g * \sin\theta$.

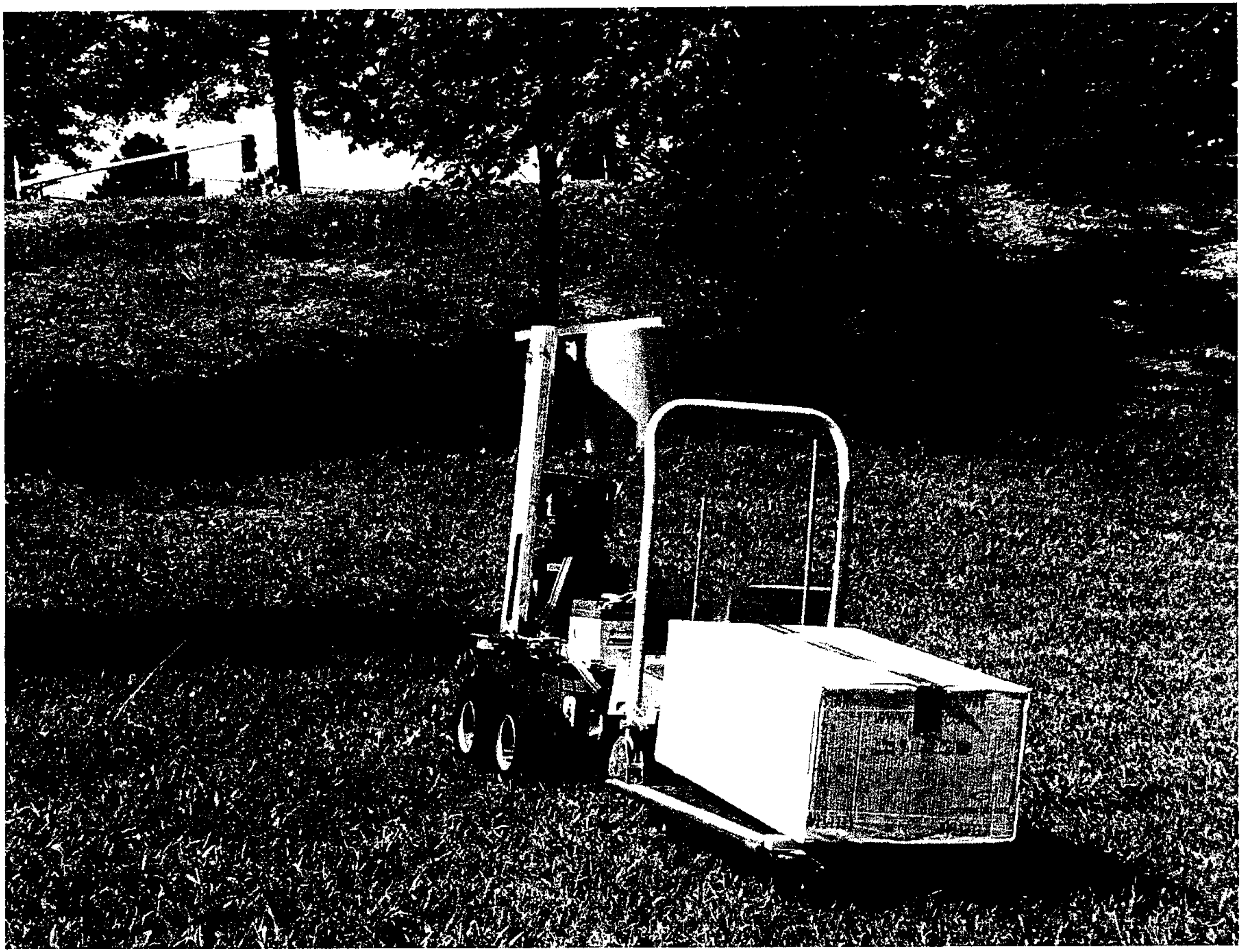


Fig. 62

Let us assume that the weight of the cargo-carrying cart is much higher than that of the robot. Then the force of friction on the wheels of the robot is $\mu_{friction} * m_{robot} * g * \cos\theta$. So, for the robot to successfully pull the cart up the incline requires that

$$\mu_{friction} * m_{robot} * g * \cos\theta > m_{cart} * g * \sin\theta.$$

In other words,

$$m_{robot} > \frac{(m_{cart} * \tan\theta)}{\mu_{friction}}$$

must hold. Given m_{robot}, the unknowns are $\mu_{friction}, m_{cart}$ and θ. These might be roughly estimated visually and from experience by the robot: m_{cart} from the size of the cargo; $\mu_{friction}$ by the color and glossiness of the incline's surface; and θ from stereo vision estimates of its depth at the top and bottom. These estimates however will have large uncertainties associated with them. Consider just one of these unknowns, m_{cart}, and define $m_{cart} = m_{vehicle} + m_{cargo}$. Based on manufacturer's specifications and known variabilities (e.g. how worn the tires are), $m_{vehicle}$ is likely to be known within an interval, such as $k \pm 10\%$ for some k. Given a sufficiently sized database of actual robot missions, m_{cargo} could be represented by a probability distribution. Therefore to calculate m_{cart} one must add two quantities one of which is an interval and the other a distribution. An easy way to do this is to substitute a second interval for the distribution, thereby making the sum easy to calculate but losing potentially valuable information. Another easy way out would be to substitute a conjectured distribution for the interval, such as a triangular distribution with mode k. But this has another problem, which is that the results of a simulation based on this model would be undependable because it uses a distribution that is merely a conjecture. The ideal approach would be to add an interval and a distribution. The problem is that Monte Carlo is not well suited to such a situation. In general it would require combining multiple Monte Carlo simulations, one for each of a random sample of values from the interval. (Well-behaved models could be handled by sampling only the endpoints of the interval.)

Lack of knowledge about the precise nature of the dependency relationships among model variables

The Monte Carlo approach may proceed straightforwardly if variables are assumed independent. However, if vaiables are not known for sure to actually be independent, resulting conclusions can be suspect. This is illustrated by the following example.

Example 4: Consider the case of a model with two uncertain variables that must be combined. Some bat populations have suffered fluctuations in population in recent years, due to such factors as pesticides in their diet of insects, other human disturbance of habitats, and perhaps other poorly understood factors. In order to estimate the number of bats of a particular species in a particular area that will be present one year from now, one can add to the current population the product of the current population and the growth rate (a negative growth rate would signify a decline in population). Neither the current population nor the growth rate is likely to be known accurately, and therefore might be better modeled using distributions than point values. Thus we would have to multiply two distributions together to get an estimate of next year's population. Furthermore, the dependency relationship between these two distributions is unknown. They could be completely independent (which could lead to eventual extinction for negative growth rates). Alternatively, they could be positively correlated. This can occur in populations that are sufficiently low to be marginally viable. In that case an increase in population can cause an increase in growth rate. Finally, it is possible for population and growth rate to be negatively correlated. This can occur for example when a population approaches the limit of the ability of the environment to support it, at which point individuals are forced into competition with each other for food and perhaps other resources, making it harder for them to survive and reproduce.

Thus, a model might specify distribution functions for population and growth rate, but the dependency relationship between the two distribution functions may be unknown. What, then, can be said about the product of population and growth rate, and hence about the population in a year? If we assumed the distributions were independent then the result of multiplying them would be some distribution. On the other hand if we assumed the distributions were completely correlated (so a higher value for one implied a correspondingly higher value for the other), or negatively correlated (a higher value for one implied a correspondingly lower value for the other) then the result in each case would again be some distinct distribution.

Since we cannot justify any particular dependency relationship in this example, the result could be any of a family of distributions, each one corresponding to some dependency relationship – whether simple or complex – between the variables. Then the family of all possible result distributions, which includes independence, full positive and negative correlation, and all other dependency relationships, may be expressed using a bounded family of distributions to represent the space within which each member of that infinitely numerous family must be (Berleant and Goodman-Strauss 1998).

The sensitivity of any conclusions to an independence assumption can be checked, to a degree, by also running a Monte Carlo simulation on the problem under the assumption that the variables are perfectly positively correlated, as well as under the assumption that some are perfectly negatively correlated with others. These different assumptions, representing extremes of possible dependencies, will lead to possibly differing conclusions (though not necessarily to extremes within the space of conclusions implied by the space of possible dependencies, see Ferson et al. 2004). This will help test the sensitivity of the conclusions to assumptions about dependency.

The trustworthiness of a Monte Carlo simulation will generally be benefited when the dependencies among the variables are known. Correlations might be known even when full details of dependencies are not. If a correlation between two variables is positive, then a relatively high sample value for one variable would typically increase the probability of a relatively high sample value of the other variable. Similarly, a negative correlation would typically increase the probability of drawing a relatively low value of the other variable. The term "typically" applies because a positive correlation can hide a tendency for some high values of one variable to occur with low values of the other, if that tendency is overcompensated by a tendency for other high values of one to occur with high values of the other (Ferson et al. 2004). Even when correlation is known and modeled, underlying details about a dependency relationship that are hidden by the crude measure of correlation could impact the validity of the model and hence dependability of the results of a Monte Carlo simulation.

1.2 How bounded families of distributions can help

We have just described how Monte Carlo simulation can be facilitated through unsupported assumptions (modeling an interval as a distribution, or assuming a dependency relationship), or discarding information (as when modeling a distribution with an interval), or kludgey 2nd-order modifications of the clean classical Monte Carlo approach. Ideally though, elegant techniques would be used that do not lead to reductions in information quality (Wang et al. 2005). The approach described next, the DEnv technique, meets that requirement.

We begin by reviewing salient features of probability distributions. Because of their familiarity, they form a convenient lead-in to a discussion of bounded families of distributions.

The probability is 0 that a sample drawn from a probability density function will be less than the lowest value in its support, and 1 that it will be no greater than the greatest value in its support (Figure 1.3). More generally, the probability that a sample will not exceed a specific value increases the value specified increases. Based on that observation, a curve that plots probability against progressively increasing given values is called a cumulative distribution function (CDF), often abbreviated simply as "distribution," Figure 1.3b.

Fig. 1.3. A probability density function (PDF) and its corresponding cumulative distribution function (CDF). A CDF describes the cumulative area under its corresponding PDF, rising to a final value of 1. The capacity of these isomorphic representations to describe uncertainty is limited, motivating more general methods.

What happens if something can be said about a density function but not enough to specify it fully? For example the mean and variance might be known, but not the detailed form of the curve. In such a situation, a family of different density functions conforms to the limited information we have about it. Almost any density function can be shifted right or left until its mean is a given value, and then stretched or compressed around the mean to adjust its variance to another given value. Such a family of curves, if many were superposed, would form a jumble and be difficult to work with. Fortunately this apparent jumble can be expressed in the more visualizable way discussed next.

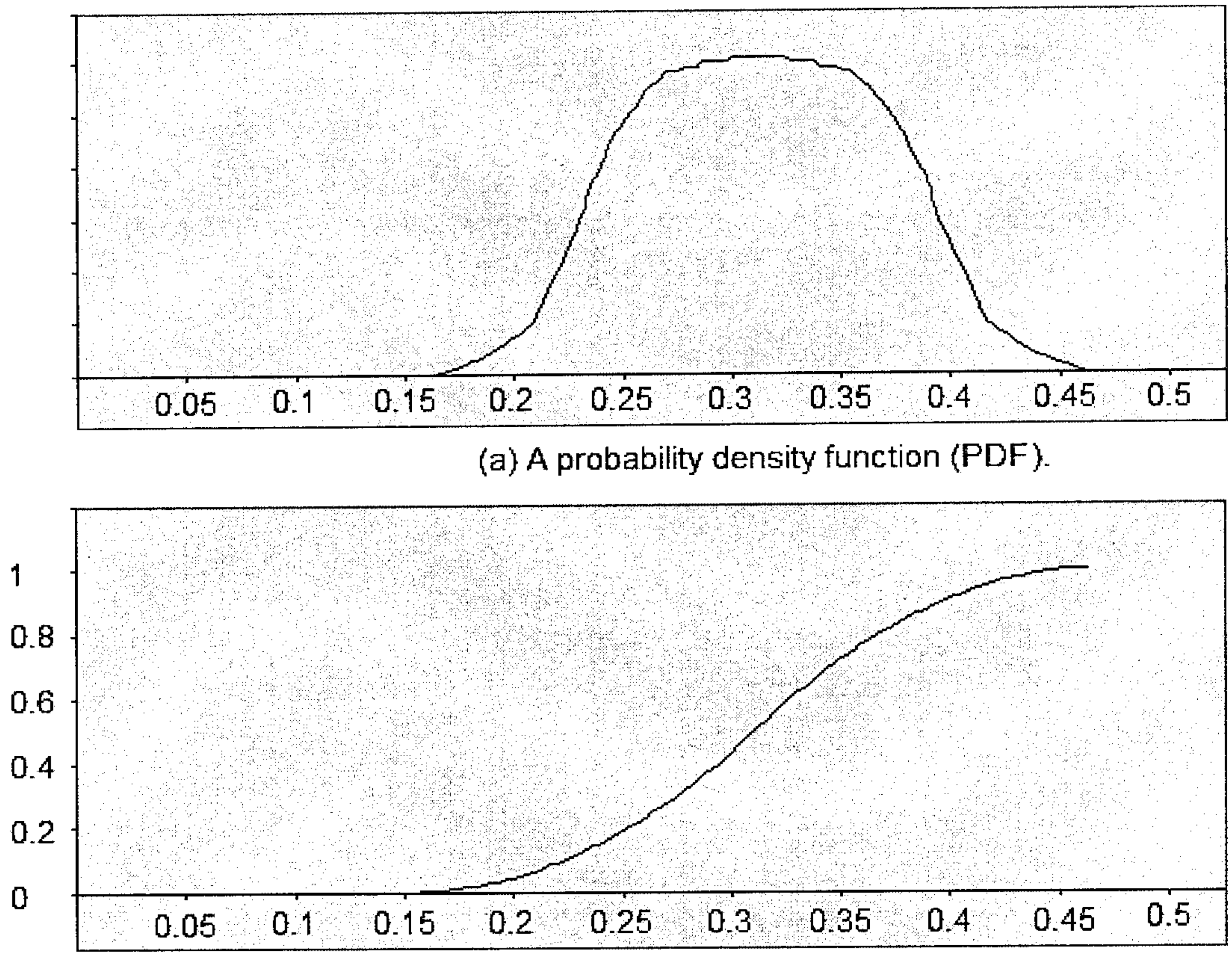
1.2.1 Bounded families of distributions

If we integrate each member of a family of density functions to get a corresponding family of distributions, it is considerably easier to visualize and work with. Figure 1.4 shows envelopes bounding a family of distributions. This family corresponds to the family of all density functions with a given mean and variance. The envelopes shown, one bounding the family on the left and one on the right, are the bounds on this family of distributions.

Fig. 1.4. Bounds around the family of cumulative distributions with mean 10 and variance 5. All such CDFs fall within these bounds, and some CDF in the family touches any given point on each bound. However, the bounding envelope curves do not themselves have mean 10 and variance 5. (The tails taper off to ±∞, not shown.)

Clearly, distribution family envelopes provide easily visualized bounds on the space through which members of the family can travel. Therefore, they also implicitly bound the corresponding family of density functions which, as noted, is not as easy to visualize directly.

Let us next show that bounded distribution families enable a general strategy for circumventing the problems of traditional Monte Carlo simulation described above. What is needed is a representation for uncertainty that can (1)



(a) A probability density function (PDF).

(b) The cumulative distribution function (CDF)
corresponding to the pdf of (a).

Fig. 13

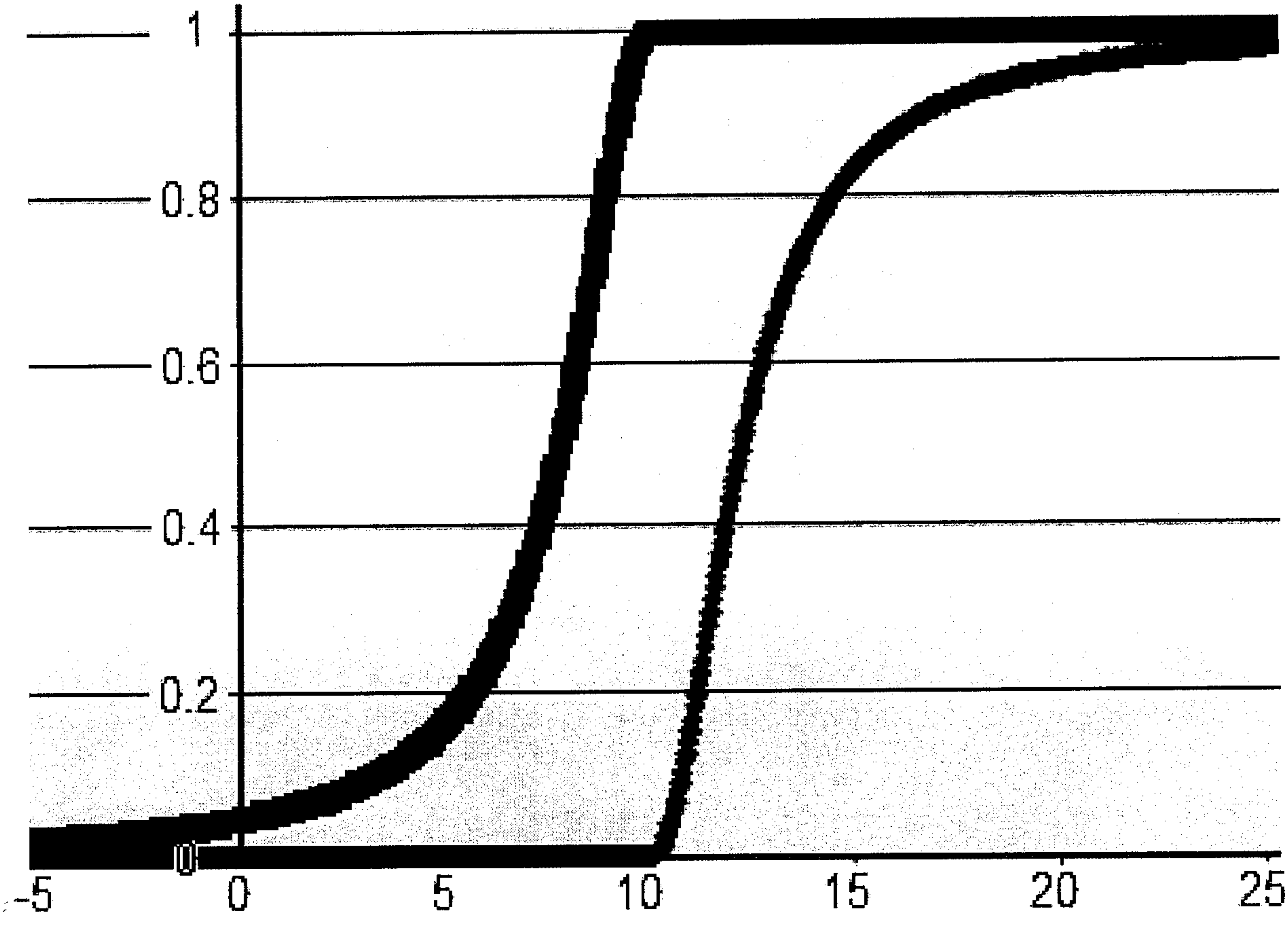


Fig.1.4

express intervals, distributions, and families of distributions, and (2) *manipulate* model variables thus represented.

1. *Expressing* intervals, distributions, and bounded families of distributions. All of these can be expressed using bounded distribution families as a unifying representation, as we see next.
   a) *Families of distributions* are described using bounds as explained above. In principle, there are families that cannot be described with bounds, for example, the family of all density functions with a single impulse and zero density everywhere else. In practice, bounding envelopes can represent the kinds of families of distributions that seem to typically arise in practice.
   b) *Distributions* are described using bounding envelopes easily, because a distribution is simply a family of distributions with one member. The appropriate bounds consist of a left envelope and a right envelope that are identical and equal to the distribution in question.
   c) *Intervals* are easily described using bounded distribution families. A variable restricted to be within an interval $[\underline{x}, \overline{x}]$ has a density function with zero density for values below $\underline{x}$ and above $\overline{x}$. Therefore any density function that integrates from 0 to 1 over the interval is in the family of distributions consistent with the interval. The extremes giving the left and right envelopes of the corresponding distribution family are therefore a density function with an impulse at $\underline{x}$ and zero density everywhere else, and a density function with an impulse at $\overline{x}$ and zero density everywhere else. See Figure 1.5.

Fig. 1.5. Bounded family of distributions whose left and right envelopes (shown with dots and dashes, respectively) represent the interval $[\underline{x}, \overline{x}]$ (with low bound $\underline{x}$ and high bound $\overline{x}$).

2. *Manipulating* model variables that may be intervals, distributions, or families of distributions. Once the variables we wish to manipulate (e.g. by adding them together, or subtracting, multiplying, dividing, or applying some other binary function to them) are all expressed as bounded families of distributions, we need no more than a method of manipulating these bounded families. In other words, the conceptual differences among intervals, distributions, and distribution families become irrelevant to the manipulation method. We address such a method next.

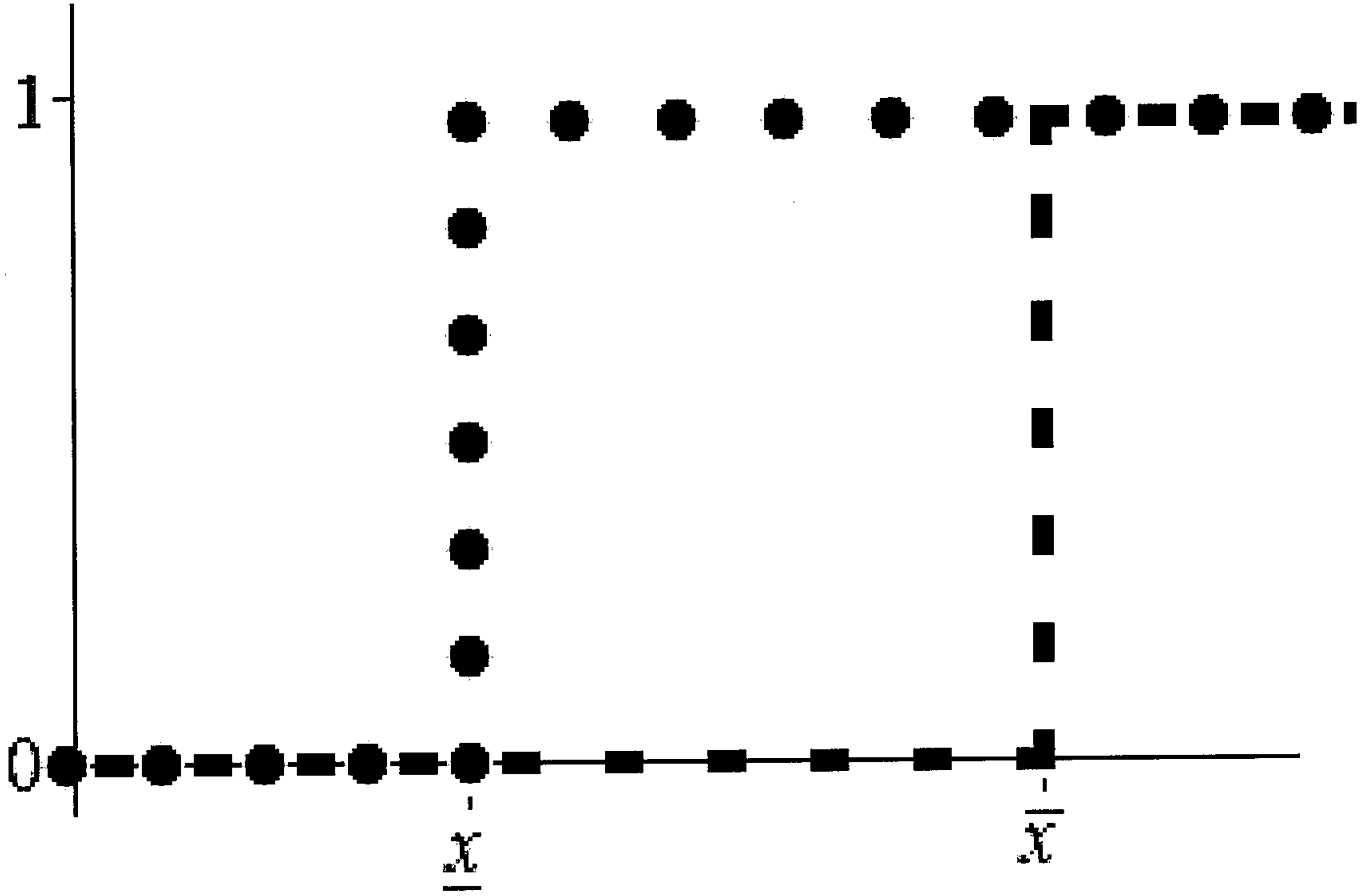


Fig. 1.5

1.3 Arithmetic operations on bounded families of distributions

We begin by showing how to apply a binary operation (e.g. addition) to variables when one is a distribution and the other is an interval. We will then extend the ideas to the other three cases of interest, one variable an interval and the other a bounded family of distributions, one a distribution and the other a bounded family, and both bounded families.

1.3.1 When one variable is a distribution and the other is an interval

Consider the case where one variable is described using a distribution function and another is less well characterized, being described only by an interval describing its range of plausible values. The presence of a variable x described by an interval typically prevents representing the sum, product, etc. of x and some distribution, as a distribution. As Figure 1.6 shows, each possible value of the interval, when combined with the distribution, leads to a distinct distribution for the output variable. Each distinct distribution is the distribution of the sum given some particular sample value from the interval. The result is a family of distributions, one for each value in the interval. This family may be bounded with envelopes.

Fig. 1.6. The distribution farthest to the left is added to the interval $[0.228, 0.421]$. The result is the bounded family of distributions on the right, of which the left and right envelopes and five example interior members are shown. In this situation none of members of the family cross each other (but in other situations, they do).

Up to this point, bounded distribution families have been illustrated – literally – graphically. But computer software for working with these families needs to represent them using numbers instead. The next section presents a method for calculating with bounded distribution families that can be implemented in computer software.

1.4 A numerical approach to computing with bounded distribution families

A suitable way to work with bounded families of distributions on computers uses sets of *intervals* associated with *probabilities*. An interval, for this purpose, is a range described by its low and high end points. The interval containing all

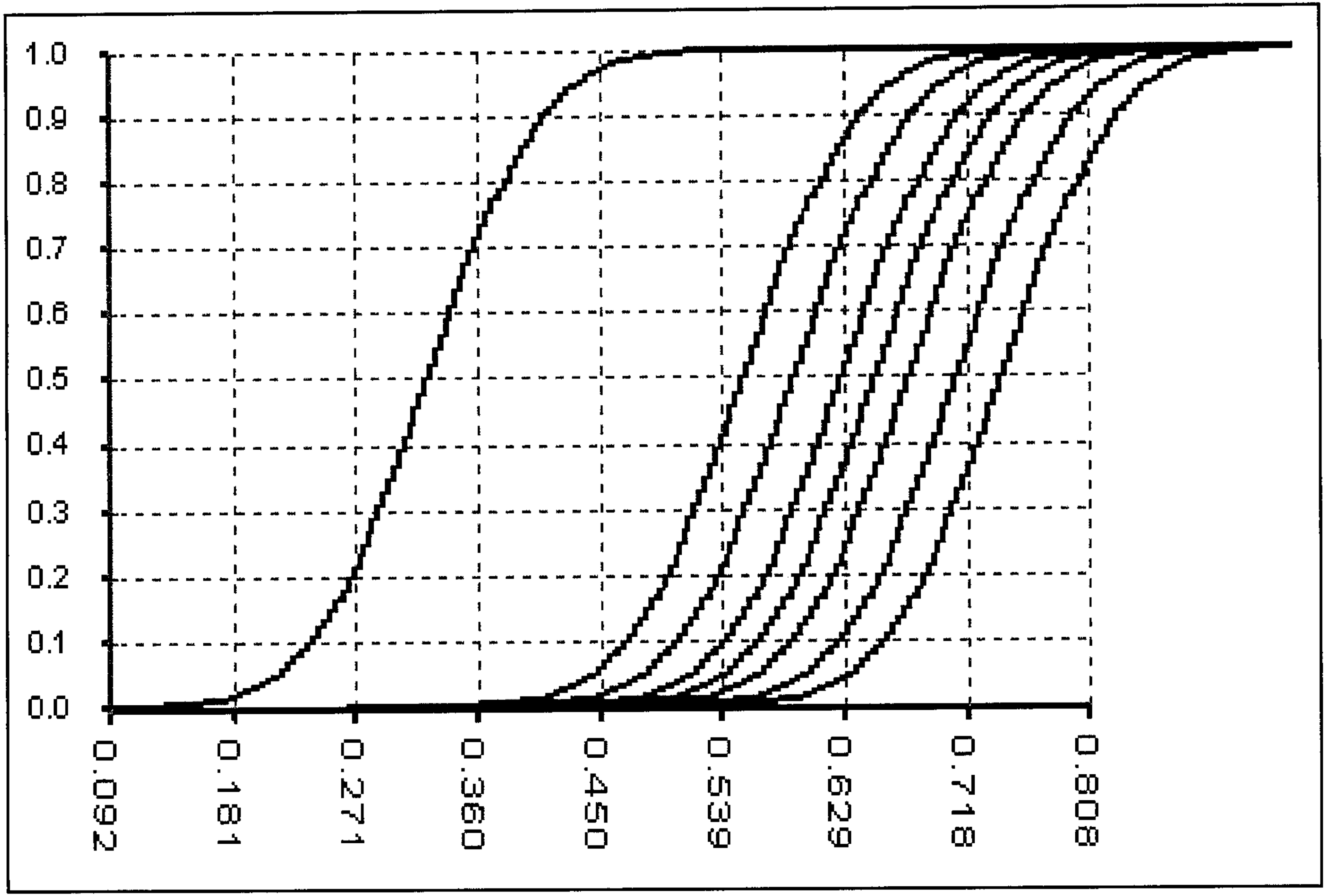

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.092
0.181
0.271
0.360
0.450
0.539
0.629
0.718
0.808

Fig. 16

numbers from 2 through 9.8 for example is written [2, 9.8]. In this approach, each interval is associated with the probability that a sample value of a random variable will belong to that interval. Graphically, a rectangle can be placed on the x-axis with its left and right sides at the low and high bounds of the interval, which has area = probability, so height = area/width. Figure 1.7 shows an example of a set of intervals with associated probabilities, its rectangles, and its bounded distribution family.

A partially specified distribution

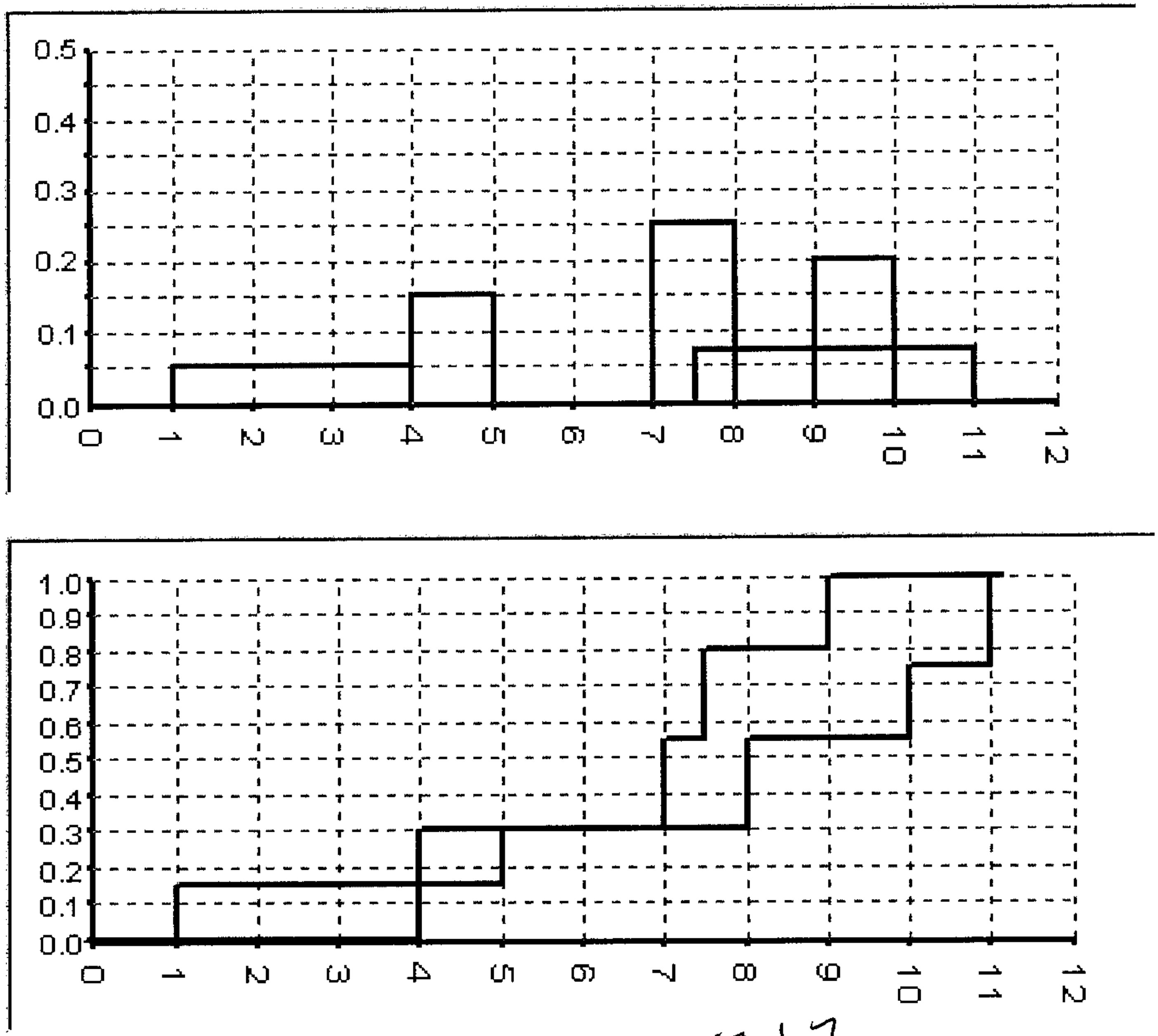
```
{p([1,4])    = 0.15
 p([4,5])    = 0.15
 p([7,8])    = 0.25
 p([7.5,11])= 0.25
 p([9,10])   = 0.2
}
```

Fig. 1.7. An under-specified distribution consisting of a set of intervals and their associated probabilities, the corresponding rectangles, and the bounded distribution family.

The rectangles are misleading in an important respect: they suggest that the distribution of probability for a given interval is uniform, because the interval and its probability are depicted using a rectangle with a flat top. In fact, no constraint on how probability is distributed within an interval is intended. At one extreme, probabilities might be concentrated as impulses at the low bounds of their intervals (i.e., at the left sides of the rectangles). Then the distribution family envelope curve will rise suddenly at the low bound of each interval (see the left staircase curve of Figure 1.7), yielding the left envelope of a bounded distribution family, which is the fastest-rising curve that is consistent with the set of intervals and their probabilities. The opposite extreme would be to concentrate the probabilities at the high bounds of their corresponding intervals. Then the cumulative curve will rise suddenly at the high bounds of the intervals yielding a staircase curve which is the righthand envelope of the bounded distribution family.

The left and right staircase shaped envelope curves are bounding in that all curves that result from distributions of probabilities within their associated intervals travel between the two envelopes, never crossing them.

For example, Figure 1.8 shows envelopes and two other distributions that are consistent with those envelopes. One distribution has three straight segments, each corresponding to a uniform distribution within one of the histogram bars in the inset. Thus, when the probability of each interval is distributed uniformly over the interval (as suggested by the flat tops of the



Other part of Fig. 1.7

histogram bars in the inset), the cumulation rises in a series of connected, nonvertical line segments between the envelopes (dark middle curve). The smooth s-curve also shown between the envelopes corresponds to some smooth density function which the histogram discretizes. Such a density function is shown superposed on the histogram.

Fig. 1.8. Left and right staircase-shaped envelopes. In general these envelopes may touch at one or more points, but they never cross. Within those bounds two CDFs are shown, one composed of three straight line segments, and one an s-curve. The inset shows rectangles arranged in a histogram and, superposed, an example of a density function that the histogram discretizes.

Showing distribution family bounds avoids problems with collections of rectangles, such as flat tops, which are misleading in seeming to suggest that probabilities are distributed uniformly over their intervals. Another potentially misleading visual characteristic of rectangle collections is that rectangles that overlap may lead to ambiguity regarding the identities of the intervals underlying them.

While showing rectangles has limitations, so does showing bounding envelopes. Different sets of intervals and their probabilities can yield the same envelopes. For example, consider the sets S_1 and S_2:

$S_1 = p([1,4]) = 0.5, p([2,3]) = 0.5$

$S_2 = p([1,3]) = 0.5, p([2,4]) = 0.5$

In both those sets, the extreme case of concentrating probabilities at the low bounds of their intervals yields two impulses, one at 1 and the other at 2. The other extreme case, concentrating the probabilities at the high bounds of the intervals, also yields two impulses, one at 3 and the other at 4. Thus the envelopes for S_1 and S_2 are identical. Yet, a random variable governed by S_1 can lead to different results than one governed by S_2. See Berleant and Zhang (2004b, section 3.1) for a fuller discussion.

1.4.1 Discretization and bounded families of distributions

Envelopes can be used for representing intervals, distributions, and bounded families of distributions. And, we have been building the case that a suitable underlying data structure for expressing a pair of envelopes is a set of intervals and their associated probabilities. To further make this case requires addressing how to express smoothly curving distributions and envelopes using these sets. Intervals have definite endpoints and, graphically, sets of them yield left and right, sharply angled staircase-like envelope curves. These are decidedly not smoothly curving. Yet sets of intervals and probabilities can

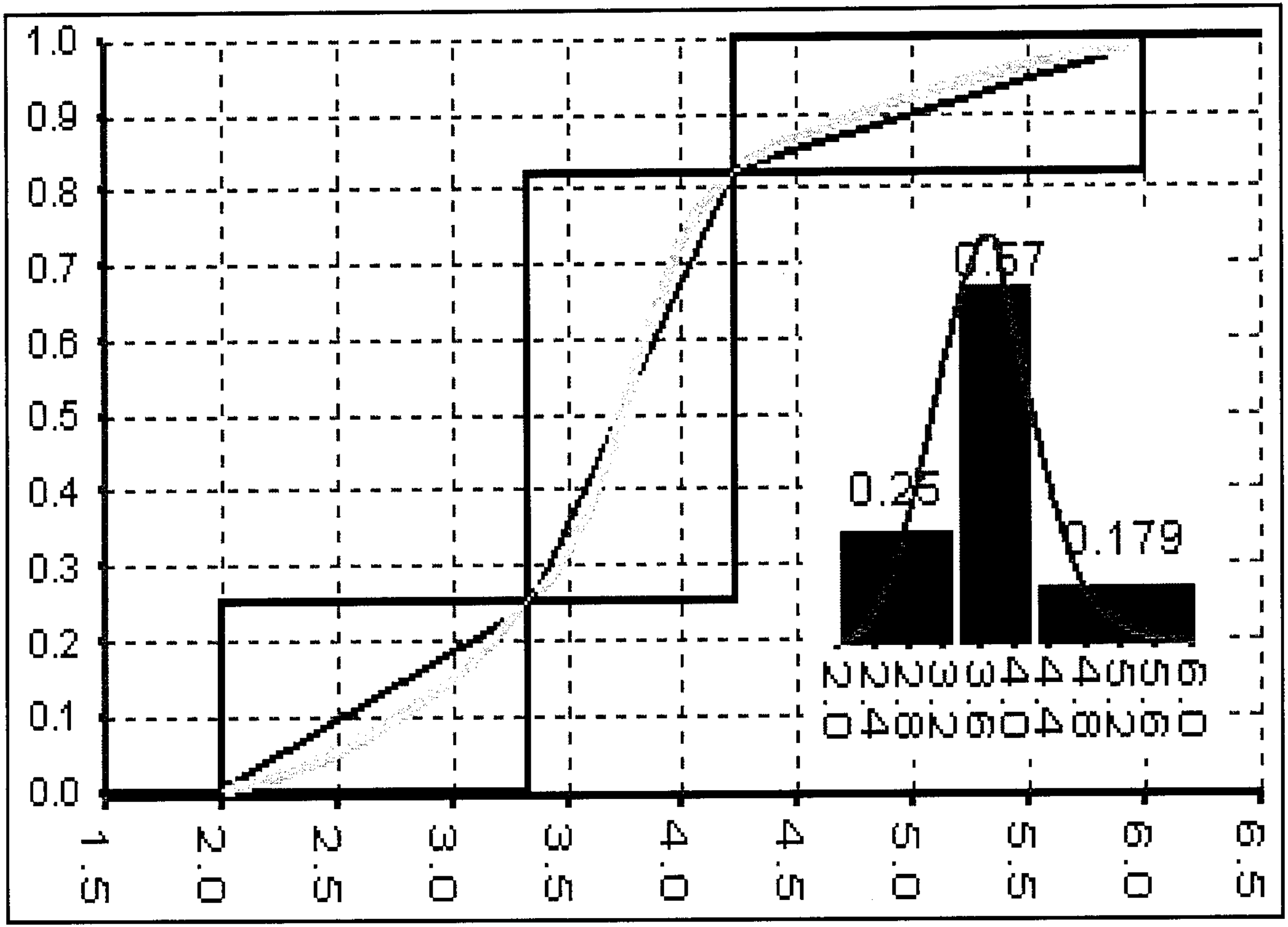

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
0.25
0.67
0.179
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0

Fig. 18

be used for representing smoothly curving envelopes as well. Figure 1.8, inset, shows a coarse discretization (rectangles forming a histogram) of a curved density function. Histograms will approximate a probability density curve better as the number of histogram bars increases and their widths decrease. In cumulative terms, a distribution will be better approximated by its enclosing staircase shaped envelopes the more steps the envelopes possess. See Figure 1.9.

Fig. 1.9. Two discretizations of the same distribution, a light-colored pair of envelopes with 4 steps each, and a dark pair with 64. Each bar of a histogram that discretizes a density function corresponds in the world of distributions (the integrals of density functions) to a box of which the north and east sides are formed by the left envelope, and the south and west sides by the right envelope.

At this point we have introduced families of distributions with three alternative representations, (1) sets of intervals and probabilities, (2) rectangles, and (3) envelopes. The sets of intervals and probabilities are the underlying, computer-friendly specification, while rectangles and envelopes are human-friendly and derivable from the sets. We have not yet shown, however, how to take two different variables, each a bounded family of distributions, and add, subtract, multiply, or divide them, or perform some other binary operation on them. This is discussed in the next section.

1.5 Computing with bounded distribution families

We introduce how to do arithmetic computations on bounded distribution families using an example with a typical structure but artificial data.

> *Example 5:* Consider the goal of finding out what can be determined about the total amount of some pesticide released into the environment worldwide. Model this total as C where $C = A + B$, A is the amount contributed by U.S. agriculture, and B is the amount contributed by all other countries. The exact values of A and B are unknown, but we assume distributions for them are available. We can discretize such distributions visually as histograms, or as left and right envelopes (similar to those in Figure 1.9), or alternatively for computational purposes as sets of intervals and their probabilities. Table 1.1 shows the description in terms of intervals and probabilities.

To compute $A + B$ to get the total amount of pesticide released, consider that for any plausible value of A, B might potentially have any value permitted by its distribution. In terms of the intervals of Table 1.1, we add each interval

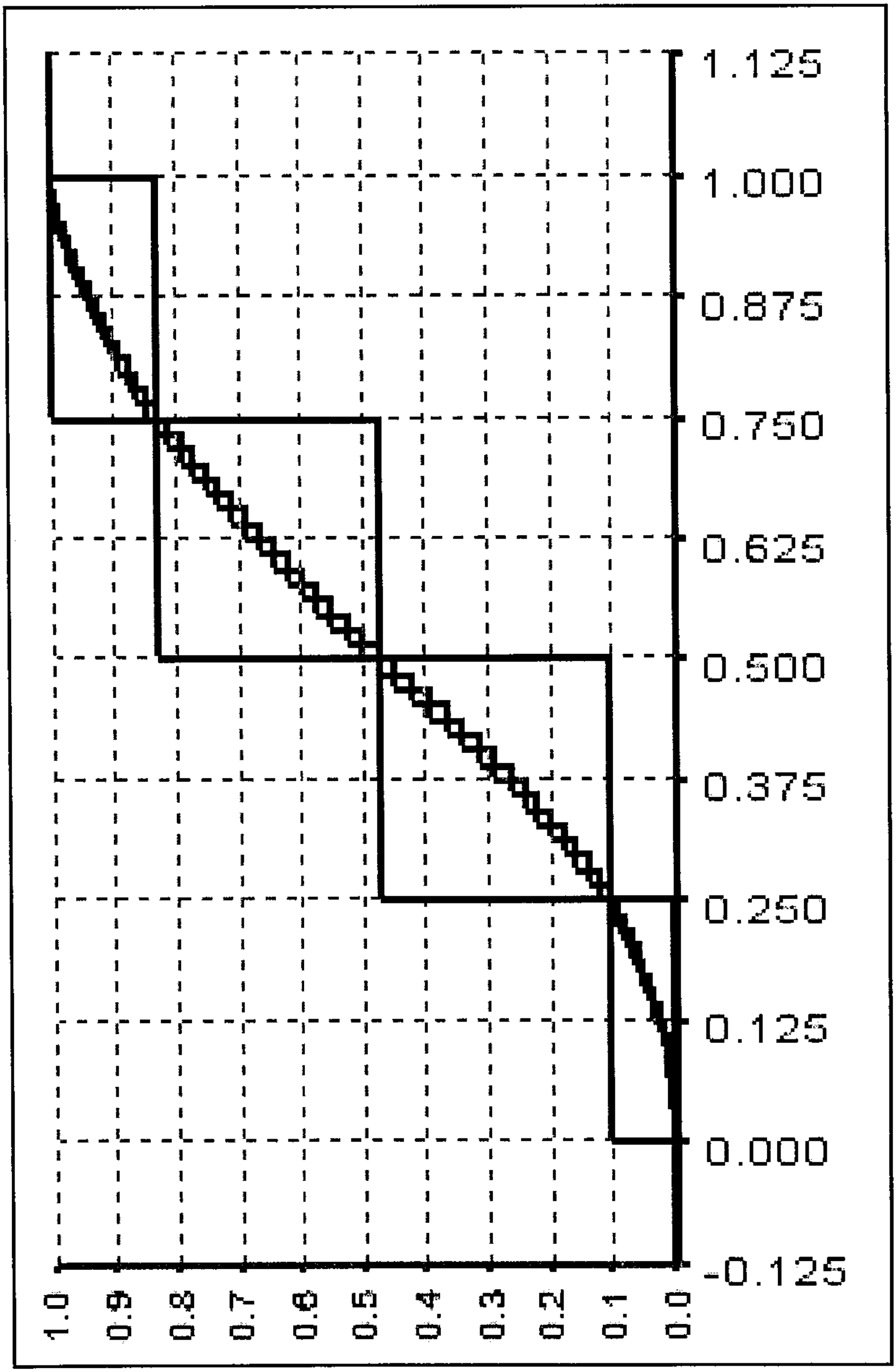

1.125
1.000
0.875
0.750
0.625
0.500
0.375
0.250
0.125
0.000
-0.125
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Fig. 19

A	B
`p(A in [10,11])=0.1`	`p(B in[5,6])=0.05`
`p(A in [11,12])=0.2`	`p(B in[6,7])=0.06`
`p(A in [12,13])=0.4`	`p(B in[7,8])=0.08`
`p(A in [13,14])=0.2`	`p(B in[8,9])=0.1`
`p(A in [14,15])=0.1`	`p(B in[9,10])=0.21`
	`p(B in[10,11])=0.21`
	`p(B in[11,12])=0.1`
	`p(B in[12,13])=0.08`
	`p(B in[13,14])=0.06`
	`p(B in[14,15])=0.05`

Table 1.1. The distribution functions describing the amounts contributed by pesticide sources A and B have been discretized and are shown symbolically as sets of intervals with associated probabilities.

in A to each interval in B to get $5 * 10 = 50$ new intervals, and calculate a probability for each of the new intervals, resulting in a set of intervals and their probabilities for $C = A + B$. Thus, if A is in $[10, 11]$, and B is in $[5, 6]$, then C would be in $[10, 11] + [5, 6] = [15, 17]$. Similarly, we can get an interval describing the value of C given A in any of its 5 intervals and B in any of its 10 intervals. See Table 1.2.

$A \quad B \rightarrow$ $\downarrow$	$[5,6]$ $p = .05$	$[6,7]$ $p = .06$	$[7,8]$ $p = .08$	$[8,9]$ $p = .1$	$[9,10]$ $p = .21$	$[10,11]$ $p = .21$	$[11,12]$ $p = .1$	$[12,13]$ $p = .08$	$[13,14]$ $p = .06$	$[14,15]$ $p = .05$
$[10,11]$ $p = .1$	$[15,17]$	$[16,18]$	$[17,19]$	$[18,20]$	$[19,21]$	$[20,22]$	$[21,23]$	$[22,24]$	$[23,25]$	$[24,26]$
$[11,12]$ $p = .2$	$[16,18]$	$[17,19]$	$[18,20]$	$[19,21]$	$[20,22]$	$[21,23]$	$[22,24]$	$[23,25]$	$[24,26]$	$[25,27]$
$[12,13]$ $p = .4$	$[17,19]$	$[18,20]$	$[19,21]$	$[20,22]$	$[21,23]$	$[22,24]$	$[23,25]$	$[24,26]$	$[25,27]$	$[26,28]$
$[13,14]$ $p = .2$	$[18,20]$	$[19,21]$	$[20,22]$	$[21,23]$	$[22,24]$	$[23,25]$	$[24,26]$	$[25,27]$	$[26,28]$	$[27,29]$
$[14,15]$ $p = .1$	$[19,21]$	$[20,22]$	$[21,23]$	$[22,24]$	$[23,25]$	$[24,26]$	$[25,27]$	$[26,28]$	$[27,29]$	$[28,30]$

Table 1.2. Intervals for A are shown down the left, for B across the top, and for $C = A + B$ in the interior cells. For example, when $A \in [10, 11]$ and $B \in [5, 6]$ then $A + B \in [15, 17]$, etc.

What about the probabilities associated with the interior cells of the table? These are not shown in the table because they vary for different dependency relationships between A and B.

We will call a table like 1.2 a *joint distribution tableau.* It shows the ranges of intervals for $C = A + B$ for all of the combinations of intervals in A and B. The probabilities associated with those intervals for C are constrained by the marginal probabilities as shown in the first row and column, but are not fully determined because there is no information available about the dependency relationship between A and B. If A and B were independent, the probabilities of the interior cells would be the product of their marginal probabilities. But A and B might not be independent. For example, heavy use of the pesticide in the US might positively correlate with heavy use elsewhere due to similar judgements of farmers worldwide. On the other hand, if overall supply was limited then heavy use in one country would limit its use elsewhere, a negative correlation. Each dependency relationship results in some distribution for C. We wish to construct the left and right envelopes around the family of all distributions plausible for C. Let us consider next a few selected values of the left and right envelopes bounding $C = A + B$, starting with the left envelope.

Left envelope

$C = 14$: There is no way for $A + B$ to be as low as 14 no matter which of the possible values of A and B occur. Indeed the lowest possible value of $A + B$ is 15, which would only occur if A and B were both at their minimum possible values of 10 and 5 respectively. Thus the left envelope is zero for $C = 14$ (and all other values of C below 15). For the same reason this is also true of the right envelope.

$C = 15$: The only way C can be 15 is if $A = 10$ and $B = 5$, which occurs only for the top left cell in the interior of the table. The probabilities in all of the interior cells in the row containing that cell sum to 0.1, because that is the marginal probability for $A \in [10, 11]$. Similarly, the probabilities associated with all of the interior cells in the *column* holding this cell must sum to 0.05 because that is the marginal probability for $B \in [5, 6]$. This puts an upper bound on the probability associated with the top left interior cell, of $\min(0.05, 0.1) = 0.05$. Some dependency relationship between A and B might be associated with such an assignment of probability to that cell, but no dependency relationship can exist for which that probability would exceed 0.05. One might guess that this upper bound of 0.05 is not achievable because of a putative need to reserve some of the 0.05 marginal probability to distribute among other cells in that column. However, simply filling in probability values in the table by hand and adjusting them by trial and error reveals that in this case, the full 0.05 probability can be allocated to the top left interior cell (Table 1.3). Later we will discuss allocating probabilities automatically.

Thus, the left envelope jumps from 0 to 0.05 at $C = 15$. This value cannot rise further until $C = 16$ because at that value for C, other interior cells associated with other ranges of A and B can contribute their probabilities to the ways in which C can be 16, as described next.

$C = 16$: The three cells whose summed probability we need to maximize in this case are the top left interior cell (call it the corner cell for now), the cell to its right, and the cell below it. As in the previous case, the leftmost interior column probabilities must add to $p(B = [5, 6]) = 0.05$. Also the 2nd column of interior cells must add up to 0.06. Both of those marginal probabilities can

$A \quad B \rightarrow$ $\downarrow$	$[5,6]$ $p = .05$	$[6,7]$ $p = .06$	$[7,8]$ $p = .08$	$[8,9]$ $p = .1$	$[9,10]$ $p = .21$	$[10,11]$ $p = .21$	$[11,12]$ $p = .1$	$[12,13]$ $p = .08$	$[13,14]$ $p = .06$	$[14,15]$ $p = .05$
$[10,11]$ $p = .1$	$[15,17]$ $p = .05$	$[16,18]$ $p = .05$	$[17,19]$ $p = 0$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$
$[11,12]$ $p = .2$	$[16,18]$ $p = 0$	$[17,19]$ $p = .01$	$[18,20]$ $p = 0.08$	$[19,21]$ $p = .1$	$[20,22]$ $p = .01$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23.25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$
$[12,13]$ $p = .4$	$[17,19]$ $p = 0$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = .2$	$[22,24]$ $p = 0.2$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$
$[13,14]$ $p = .2$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = .1$	$[25,27]$ $p = .08$	$[26,28]$ $p = .02$	$[27,29]$ $p = 0$
$[14,15]$ $p = .1$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0.01$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$	$[27,29]$ $p = .04$	$[28,30]$ $p = .05$

Table 1.3. Joint distribution tableau for $C = A + B$. Each interior cell shows a range and probability for C associated with an interval and probability for B at the head of its column, and for A at the head of its row. Many other probability asignments are also possible, corresponding to different dependencies between A and B.

be distributed among just those 3 cells, leaving other interior cells in the two leftmost columns with zero probabilities. This maximizes the summed probability of those 3 cells at 0.11. Manual inspection of the problem is one way to reveal that such an allocation is possible (Table 1.4). This maximum probability of 0.11 applies for values of C from 16 up to 17, at which point other interior cells can contribute their probabilities to the ways in which C can be 17.

$A \quad B \rightarrow$ $\downarrow$	$[5,6]$ $p = .05$	$[6,7]$ $p = .06$	$[7,8]$ $p = .08$	$[8,9]$ $p = .1$	$[9,10]$ $p = .21$	$[10,11]$ $p = .21$	$[11,12]$ $p = .1$	$[12,13]$ $p = .08$	$[13,14]$ $p = .06$	$[14,15]$ $p = .05$
$[10,11]$ $p = .1$	$[15,17]$ $p = \mathbf{.04}$	$[16,18]$ $p = \mathbf{.06}$	$[17,19]$ $p = 0$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$
$[11,12]$ $p = .2$	$[16,18]$ $p = \mathbf{.01}$	$[17,19]$ $p = 0$	$[18,20]$ $p = .08$	$[19,21]$ $p = .1$	$[20,22]$ $p = .01$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$
$[12,13]$ $p = .4$	$[17,19]$ $p = 0$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = .21$	$[23,25]$ $p = .1$	$[24,26]$ $p = .08$	$[25,27]$ $p = .01$	$[26,28]$ $p = 0$
$[13,14]$ $p = .2$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = .2$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$	$[27,29]$ $p = 0$
$[14,15]$ $p = .1$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$	$[27,29]$ $p = .05$	$[28,30]$ $p = .05$

Table 1.4. Three interior cells contribute to the cumulative probability $p(C \leq 16)$. These are clustered in the upper left.

$C = 17$: To get the value of the left envelope at this value, we need to maximize the sum of the probabilities in six cells (clustered in the top left area of Table 1.5) whose intervals contain any values equal to 17 or less. That table manages to allocate probabilities so that all of the probabilities in the first three interior columns are allocated within those six cells. The sum of the probabilities associated with those cells, 0.19, is maximal because any more probability would violate the constraints imposed by the marginal values of probability for B shown along the top of the table. This maximized probability applies over $17 \leq C < 18$.

A $B \rightarrow$ $\downarrow$	$[5,6]$ $p = .05$	$[6,7]$ $p = .06$	$[7,8]$ $p = .08$	$[8,9]$ $p = .1$	$[9,10]$ $p = .21$	$[10,11]$ $p = .21$	$[11,12]$ $p = .1$	$[12,13]$ $p = .08$	$[13,14]$ $p = .06$	$[14,15]$ $p = .05$
$[10,11]$ $p = .1$	$[15,17]$ $p = 0$	$[16,18]$ $p = 0$	$[17,19]$ $\boldsymbol{p = .08}$	$[18,20]$ $p = .02$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$
$[11,12]$ $p = .2$	$[16,18]$ $p = 0$	$[17,19]$ $\boldsymbol{p = .06}$	$[18,20]$ $p = 0$	$[19,21]$ $p = .08$	$[20,22]$ $p = .06$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$
$[12,13]$ $p = .4$	$[17,19]$ $\boldsymbol{p = .05}$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = .15$	$[22,24]$ $p = .2$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$
$[13,14]$ $p = .2$	$[18,20]$ $p = 0$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = .01$	$[24,26]$ $p = .1$	$[25,27]$ $p = .08$	$[26,28]$ $p = .01$	$[27,29]$ $p = 0$
$[14,15]$ $p = .1$	$[19,21]$ $p = 0$	$[20,22]$ $p = 0$	$[21,23]$ $p = 0$	$[22,24]$ $p = 0$	$[23,25]$ $p = 0$	$[24,26]$ $p = 0$	$[25,27]$ $p = 0$	$[26,28]$ $p = 0$	$[27,29]$ $p = .05$	$[28,30]$ $p = .05$

Table 1.5. The cumulative probability for $C \leq 17$ is maximized by assigning probabilities to interior cells as shown. The probabilities contributing to the sum are bolded. The full marginal probabilities of those columns may be assigned to the three interior cells holding the interval $[17,19]$, so their summed probability of $0.05 + 0.06 + 0.08 = 0.19$ is the maximum possible cumulation at $C = 17$.

We can continue to work out the values of the left envelope for higher and higher values of C, but a nïaive, pencil-and-paper approach gets unwieldy for mid-range values of C. Furthermore, computers do not use pencil and paper, but require a well-defined procedure. Before discussing such a procedure, however, let us get a start on the right envelope, illustrating its nature as the dual of the left.

Right envelope

$C = 21 - \epsilon$: To get the cumulative probability defining a y-axis value of a point on the right bounding envelope of C, we must *minimize* the sum of the probabilities associated with interior cells whose interval high bounds are below C and which therefore *must* contribute all their probability to the cumulation. Every other interior cell holds an interval with a high bound above C, and so either *cannot* contribute probability to C (if its low bound is also above C), or *might not* contribute probability to C (because its probability could be concentrated at its high bound which is above C even though its low bound is below C).

Table 1.6 shows an allocation of probabilities to interior cells that minimizes the sum of probabilities in cells whose intervals have high bounds *below* 21 and whose probabilities must therefore contribute to the accumulated probability at $C = 21 - \epsilon$ for small enough ϵ. In this case, the summed probability can be as low as zero, as the table illustrates.

$C = 21$: The minimum cumulative probability at $C = 21$ consists of the minimum possible sum of the probabilities of cells whose intervals have high bounds of 21 or less. This value is 0.05, because the set of cells whose summed probabilities is to be minimized includes the entire first column (which is constrained by the marginal probabilities of B to contain a total probability of 0.05 within its cells), and the table can be arranged so no other probability is allocated within the set of cells in question. Table 1.7 illustrates a way to do this.

A $B \rightarrow$ $\downarrow$	$[5,6]$ $p=.05$	$[6,7]$ $p=.06$	$[7,8]$ $p=.08$	$[8,9]$ $p=.1$	$[9,10]$ $p=.21$	$[10,11]$ $p=.21$	$[11,12]$ $p=.1$	$[12,13]$ $p=.08$	$[13,14]$ $p=.06$	$[14,15]$ $p=.05$
$[10,11]$ $p=.1$	$[15,17]$ $p=0$	$[16,18]$ $p=0$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=.01$	$[20,22]$ $p=.01$	$[21,23]$ $p=0$	$[22,24]$ $p=.08$	$[23,25]$ $p=0$	$[24,26]$ $p=0$
$[11,12]$ $p=.2$	$[16,18]$ $p=0$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=.1$	$[20,22]$ $p=0$	$[21,23]$ $p=0$	$[22,24]$ $p=.1$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$
$[12,13]$ $p=.4$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=0$	$[20,22]$ $p=0$	$[21,23]$ $p=.2$	$[22,24]$ $p=.2$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$	$[26,28]$ $p=0$
$[13,14]$ $p=.2$	$[18,20]$ $p=0$	$[19,21]$ $p=.06$	$[20,22]$ $p=.08$	$[21,23]$ $p=0$	$[22,24]$ $p=0$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$	$[26,28]$ $p=.06$	$[27,29]$ $p=0$
$[14,15]$ $p=.1$	$[19,21]$ $p=.05$	$[20,22]$ $p=0$	$[21,23]$ $p=0$	$[22,24]$ $p=0$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$	$[26,28]$ $p=0$	$[27,29]$ $p=0$	$[28,30]$ $p=.05$

Table 1.6. An allocation of probabilities to interior cells that minimizes the accumulated probability at $C = 21 - \epsilon$. The relevant cells are those with intervals whose high bounds are below 21. This comprises 10 cells clustered in the upper left of the table. All of these can contain 0 probability while maintaining consistency with the marginal probabilities for A and B, so the minimum summed probability is zero.

A $B \rightarrow$ $\downarrow$	$[5,6]$ $p=.05$	$[6,7]$ $p=.06$	$[7,8]$ $p=.08$	$[8,9]$ $p=.1$	$[9,10]$ $p=.21$	$[10,11]$ $p=.21$	$[11,12]$ $p=.1$	$[12,13]$ $p=.08$	$[13,14]$ $p=.06$	$[14,15]$ $p=.05$
$[10,11]$ $p=.1$	$[15,17]$ $p=.05$	$[16,18]$ $p=0$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=0$	$[20,22]$ $p=0$	$[21,23]$ $p=0$	$[22,24]$ $p=0$	$[23,25]$ $p=0$	$[24,26]$ $p=.05$
$[11,12]$ $p=.2$	$[16,18]$ $p=0$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=0$	$[20,22]$ $p=.2$	$[21,23]$ $p=0$	$[22,24]$ $p=0$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$
$[12,13]$ $p=.4$	$[17,19]$ $p=0$	$[18,20]$ $p=0$	$[19,21]$ $p=0$	$[20,22]$ $p=.1$	$[21,23]$ $p=0$	$[22,24]$ $p=.2$	$[23,25]$ $p=.1$	$[24,26]$ $p=0$	$[25,27]$ $p=0$	$[26,28]$ $p=0$
$[13,14]$ $p=.2$	$[18,20]$ $p=0$	$[19,21]$ $p=0$	$[20,22]$ $p=.04$	$[21,23]$ $p=0$	$[22,24]$ $p=.01$	$[23,25]$ $p=.01$	$[24,26]$ $p=0$	$[25,27]$ $p=.08$	$[26,28]$ $p=.06$	$[27,29]$ $p=0$
$[14,15]$ $p=.1$	$[19,21]$ $p=0$	$[20,22]$ $p=.06$	$[21,23]$ $p=.04$	$[22,24]$ $p=0$	$[23,25]$ $p=0$	$[24,26]$ $p=0$	$[25,27]$ $p=0$	$[26,28]$ $p=0$	$[27,29]$ $p=0$	$[28,30]$ $p=0$

Table 1.7. Minimized cumulative probability for $C = 21$. This requires minimizing the summed probabilities of 15 interior cells in the upper left region of the table. These are the cells containing intervals with high bounds at or below 21.

This manual process of minimizing cumulated probability for different values of C (for the right bounding curve) and maximizing it (for the left), if continued, can produce the complete left and right envelopes. However, a method that can be done by computer is desirable. Such a method is described next.

1.6 Finding points on the bounding envelopes with linear programming

As the previous section explained, to find the y-axis probability value of a point on the left or right envelope for a given x-axis value, we must maximize or minimize the sum of the probabilities of some subset of the interior cells in a joint distribution tableau. The probabilities in such tables express a kind of discretized joint probability distribution of two random variables. The marginals of these tables constrain how the probabilities of the interior cells can be allocated during the process of maximizing or minimizing a sum of a subset of them. Specifically, the marginal probabilities impose a value on the sum of the probabilities of the interior cells in each *column*, as well as on the sum of the probabilities of the interior cells in each *row*. These marginal values are givens (see tables of previous section).

The preceding paragraph summarizes the need to maximize or minimize a sum given other, constant sums. This type of situation lends itself to **linear programming**, a widely used technique. Numerous software packages, commercial and public domain, exist for solving linear programming problems. Computer program listings for this are even printed in books. Therefore rather than describe linear programming algorithms, we instead show how to set up a linear programming problem whose solution is a maximized value giving the y coordinate of a left or right envelope for some x-axis value. We can assume maximization because when the objective is to minimize the summed probabilities of some interior cells, we can simply maximize the sum of the other interior cells, and subtract that value from 1.

The desired linear programming problem consists of the constraints, and the sum to be maximized (the objective function, in linear programming terminology). For illustration, consider a simpler joint distribution tableau than the one used in the previous section (Table 1.8).

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p =$
$Y \in [3,4]$ $p = \frac{1}{2}$	$XY \in [3,8]$ $p =$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p = \frac{1}{2}$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.8. A joint distribution tableau showing marginals X and Y and interior cells showing intervals for product XY. Probabilities for the interior cells are left blank because they are not fully determined.

In Table $^{1.8}$ the probabilities in the interior cells (which spread out from the northeast corner of the table) are left out because they depend on the dependency relationship between X and Y. Thus they are variable, though constrained to some degree by the marginal probabilities shown on the left and along the bottom. Linear programming can identify specific probabilities for those interior cells that are (1) consistent with the marginal constraints, and (2) maximize the summed probabilities of any given subset of interior cells.

To solve maximization problems such as these by linear programming, one initializes by assigning feasible values to the variables, which in this case are the interior cell probabilities. These values serve as a starting point from which the linear programming process will automatically find an optimal (maximizing) allocation of probabilities. An initialization method is illustrated next on the joint distribution tableau of Table 1.8.

1. Identify the row with the highest marginal probability, the column with the highest marginal probability, and the interior cell at the intersection of that row and column. The interval in this cell is emphasized in Table 1.9. (In this case both columns have the same marginal probability, so the first one was chosen.)

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p =$
$Y \in [3,4]$ $p = \frac{1}{2}$	$XY \in [\mathbf{3},\mathbf{8}]$ $p =$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p = \frac{1}{2}$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.9. $XY \in [3,8]$ is chosen as the location for an initial probability assignment.

2. Assign to the identified cell the maximum probability consistent with the row and column marginal constraints affecting it. This is the lesser of the row and column marginal probabilities. See Table 1.10.

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p =$
$Y \in [3,4]$ $p = \frac{1}{2}$	$XY \in [\mathbf{3},\mathbf{8}]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p = \frac{1}{2}$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.10. An initial probability assignment is made to the interior cell holding interval $[3,8]$.

3. For bookkeeping purposes, subtract the probability just assigned from both corresponding marginal probabilities (Table 1.11).
4. Repeat step 1: identify a row with the highest marginal probability designation, a column with the highest marginal probability, and the cell at the intersection of that row and column. This cell is the one holding the interval $[8,20]$ in Table 1.11.
5. Repeat step 2: assign to the newly identified cell the maximum probability consistent with the row and column constraints affecting it. This is the lesser of the row and column probabilities. See Table 1.12.
6. Repeat step 3 on the table as most recently modified: to keep track of marginal probability that still needs to be allocated to interior cells, sub-

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p =$	$XY \in \mathbf{[8,20]}$ $p =$
$Y \in [3,4]$ $p' = \mathbf{0}$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = \mathbf{0}$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.11. The probability assigned to the interior cell holding $[3,8]$ is subtracted from the contributing marginal probabilities, which are now labeled p' instead of p to indicate they have been modified. Then the cell holding $[8,20]$ is chosen as the next one to allocate an initial probability to.

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p =$	$XY \in \mathbf{[8,20]}$ $p = \mathbf{\frac{1}{4}}$
$Y \in [3,4]$ $p' = 0$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = 0$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.12. The cell holding interval $[8,20]$ has its probability assigned a value as high as is consistent with its marginal probabilities.

tract the probability just assigned to an interior cell from the corresponding marginal probabilities (Table 1.13).

$Y \in [4,5]$ $p' = \mathbf{0}$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p = \frac{1}{4}$
$Y \in [3,4]$ $p' = 0$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in \mathbf{[4,12]}$ $p =$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = 0$	$X \in [2,4]$ $p' = \mathbf{\frac{1}{4}}$

Table 1.13. The allocated probability is subtracted from the relevant marginal cell probabilities, whose remaining unallocated probabilities are designated p'. Then the next cell, holding interval $[4,12]$, is chosen.

7. Repeat step 1 on the current table: identify a row with the highest amount of as-yet unallocated marginal probability of the rows, a column with the highest amount, and the cell at the intersection of that row and column. This cell holds interval $[4,12]$ in Table 1.13.

8. Repeat step 2: assign to the just-identified cell the maximum probability consistent with the row and column constraints affecting it. This will be the lesser of its row and column unallocated marginal probabilities. See Table 1.14.

$Y \in [4,5]$ $p' = 0$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p = \frac{1}{4}$
$Y \in [3,4]$ $p' = 0$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p = \frac{1}{4}$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = 0$	$X \in [2,4]$ $p' = \frac{1}{4}$

Table 1.14. An initial probability is assigned to the cell holding interval $[4, 12]$.

9. Repeat step 3: subtract the initial probability assigned the cell from its corresponding marginal probabilities. See Table 1.15.

$Y \in [4,5]$ $p' = 0$	$XY \in [4,10]$ $p =$	$XY \in [8,20]$ $p = \frac{1}{4}$
$Y \in [3,4]$ $p' = 0$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p =$
$Y \in [2,3]$ $p' = 0$	$XY \in [2,6]$ $p =$	$XY \in [4,12]$ $p = \frac{1}{4}$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = 0$	$X \in [2,4]$ $p'' = 0$

Table 1.15. The probability assigned to an interior cell is subtracted from the relevant marginals, leaving all zeroes in the margins. Key: the number of apostrophes (p, p', or p'') reflects how many times an original marginal probability has been decremented.

10. All marginal probability numbers are now 0, indicating that no marginal probability remains to be allocated to interior cells. Therefore any interior cells not yet assigned an initial probability must be assigned 0 (Table 1.16).
11. Since the interior cells are now initialized appropriately, the marginal probability designations used for bookkeeping purposes are no longer needed, and can be replaced with the actual marginal probability values that were originally present. This results in Table 1.17, which serves as input to the linear programming problem.

Table 1.17 also gives distinctive subscripts to the interior cell probabilities so that they can be referred to individually. From this table, linear programming will find the best allocation of marginal probabilities over interior

$Y \in [4,5]$ $p' = 0$	$XY \in [4,10]$ $p = 0$	$XY \in [8,20]$ $p = \frac{1}{4}$
$Y \in [3,4]$ $p' = 0$	$XY \in [3,8]$ $p = \frac{1}{2}$	$XY \in [6,16]$ $p = 0$
$Y \in [2,3]$ $p' = 0$	$XY \in [2,6]$ $p = 0$	$XY \in [4,12]$ $p = \frac{1}{4}$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p' = 0$	$X \in [2,4]$ $p'' = 0$

Table 1.16. The interior cells are now fully initialized and ready for a linear programming process to modify them to an optimal set of assignments that maximizes.

$Y \in [4,5]$ $p = \frac{1}{4}$	$XY \in [4,10]$ $p_{11} = 0$	$XY \in [8,20]$ $p_{21} = \frac{1}{4}$
$Y \in [3,4]$ $p = \frac{1}{2}$	$XY \in [3,8]$ $p_{12} = \frac{1}{2}$	$XY \in [6,16]$ $p_{22} = \frac{1}{4}$
$Y \in [2,3]$ $p = \frac{1}{4}$	$XY \in [2,6]$ $p_{13} = 0$	$XY \in [4,12]$ $p_{23} = \frac{1}{4}$
$Y \Uparrow X \Rightarrow$	$X \in [1,2]$ $p = \frac{1}{2}$	$X \in [2,4]$ $p = \frac{1}{2}$

Table 1.17. The interior cells have been initialized and the table is ready for linear programming to be applied.

cells, which is the one with the maximum value possible for the sum of the probabilities of a designated subset of the interior cells. The linear programming problem takes as input all of the row and column constraints, plus the optimization (or "objective") function, which is the sum of the interior cell probabilities to be maximized. For Table 1.17, this input is shown in Table 1.18.

Value	p_{11}	p_{21}	p_{12}	p_{22}	p_{13}	p_{23}
1/4	1	1	0	0	0	0
1/2	0	0	1	1	0	0
1/4	0	0	0	0	1	1
1/2	1	0	1	0	1	0
1/2	0	1	0	1	0	1
3/4	1	0	1	0	1	1

Table 1.18. The five constraints (3 row + 2 column) from Table 9.17 are shown in the first five rows, followed in the last row by the optimization function for maximizing the sum of the probabilities of the four interior cells whose interval low bound is below 6. The 1's and 0's are coefficients of the p_{ij} from Table 9.17.

The linear programming process will take a chart like Table 1.18 and find values for the various p_{ij} that maximize the number at the bottom of the Value column. In the case shown, this is initially 3/4, the sum of the probabilities whose associated intervals have low bounds below six. Note that the initial values of the probabilities shown in Table 1.17 determine the initial value of 3/4 for the optimization function. The 1s in Table 1.18 are coefficients that designate which probabilities are governed by which constraints. Thus, the first row says that $1 * p_{11} + 1 * p_{21} + 0 * p_{12} + 0 * p_{22} + 0 * p_{13} + 0 * p_{23} = \frac{1}{4}$ or, equivalently, $p_{11} + p_{21} = \frac{1}{4}$. This is a constraint stated by the top row of Table 1.17, as are the next four rows. The remaining, last row is not a constraint but rather the optimization equation. Hence the value of 3/4 is not fixed, as it would be for a constraint, but can vary. Normally it does vary, as the linear programming alorithm tries to maximize it.

1.7 Conclusion

An introduction to the DEnv approach has been presented at the tutorial level. More advanced features are available, and a considerable amount of related work by others has appeared. Regarding advanced features, one is the use of correlation between two random variables to supplement the basic row and column constraints imposed by a joint distribution tableau. This is described in detail in Berleant and Zhang (2004c). A slightly less general, but more accessible, discussion of correlation along with an application to reliability of 2-component systems appears in Berleant and Zhang (2004a). Many joint distributions encountered in practice are unimodal. Unimodality constraints in the DEnv approach and its software implementation are discussed by Zhang and Berleant (2005). A more theoretical discussion produced by the Kreinovich lab appears in Berleant et al. (2007). A tool and related documentation is available for download at http://ifsc.ualr.edu/jdberleant/statool/index.htm.

Work on bounded families of distributions has experienced a surge of interest in recent years. Considerable work has appeared in the biannual International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA) sponsored by the epynomous society (http://sipta.org/). The focus of the biannual Workshop on Reliable Engineering Computing (http://www.gtsav.gatech.edu/workshop/rec08, .../rec06, and .../rec04) is even more apropos. The most closely related and coherent compendium of work is still Helton and Oberkampf (2004), a collection of papers all focusing on the same set of challenge problems concerning system response under uncertainty. Various alternatives to the DEnv algorithm are explored in the context of this set of problems. Also closely related are reports by Ferson et al. (2004, 2002). Although well known in the field, Kuznetsov (1991) is unfortunately currently unavailable in English. Other books of particular interest include Fellin et al. (2005), Halpern (2003), Manski (2003), Walley (1991).

References

1. D. Berleant and G. T. Anderson, "Decision-making under severe uncertainty for autonomous mobile robots," Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 7-10, 2007, Montral.
2. D. Berleant and H. Cheng, "A software tool for automatically verified operations on intervals and probability distributions," Reliable Computing 4 (1) (1998), pp. 71-82.
3. D. Berleant 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. D. Berleant, O. Kosheleva, V. Kreinovich, and H. T. Nguyen, "Unimodality, independence lead to NP-hardness of interval probability problems," Reliable Computing, 13 (3) (2007), pp. 261-282.
5. (a) D. Berleant and J. Zhang, "Bounding the times to failure of 2-component systems, IEEE Transactions on Reliability, 53 (4) (Dec. 2004), pp. 542-550.
6. (b) D. Berleant and J. Zhang, "Representation and problem solving with Distribution Envelope Determination (DEnv)," Reliability Engineering and System Safety, 85 (1-3) (2004), pp. 153-168.
7. (c) D. Berleant and J. Zhang, "Using Pearson correlation to improve envelopes around the distributions of functions, Reliable Computing, 10 (2) (2004), pp. 139-161.
8. W. Fellin, H. Lessmann, M. Oberguggenberger, R. Vieider, *Analyzing Uncertainty in Civil Engineering,* Springer-Verlag, Berlin 2005.
9. S. Ferson, J. Hajagos, D. Berleant, J. Zhang, W. T. Tucker, L. Ginzburg, and W. Oberkampf, Dependence in Dempster-Shafer theory and probability bounds analysis, Technical Report SAND2004-3072, Sandia National Laboratory, October 2004.
10. S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz, Constructing probability boxes and Dempster-Shafer structures, Sandia National Laboratories, Report SAND2002-4015, January 2003.
11. J. Y. Halpern, *Reasoning about uncertainty,* MIT Press, 2003.
12. J. C. Helton and W.L. Oberkampf, eds., Special volume on alternative representations of epistemic uncertainty, Reliability Engineering and System Safety, 85 (1-3) (July-Sept. 2004), pp. 1-369.
13. A. Kolmogoroff, Confidence limits for an unknown distribution function, Annals of Mathematical Statistics, 12(4)(1941), pp. 461-463.
14. V. Kuznetsov, Interval statistical models (in Russian), Radio i Svyaz, Moscow, 1991.
15. C. Manski, Partial identification of probability distributions, Springer-Verlag, New York, 2003.
16. P. Walley, Statistical reasoning with imprecise probabilities, Chapman & Hall, N.Y., 1991.
17. R. Wang, E. Pierce, S. Madnick, and C. Fisher, Information Quality, M. E. Sharpe, 2005.
18. J. Zhang and D. Berleant, "Arithmetic on random variables: squeezing the envelopes with new joint distribution constraints, Proceedings of the Fourth International Symposium On Imprecise Probabilities and Their Applications (ISIPTA 05), Carnegie Mellon University, Pittsburgh, July 20-23, 2005, pp. 416-422.