

1 **MYTHS ABOUT CORRELATIONS AND DEPENDENCIES**
2 **AND THEIR IMPLICATIONS FOR RISK ANALYSIS**

3
4
5
6

Scott Ferson¹, Roger Nelsen², Janos Hajagos^{1,3}, Daniel Berleant⁴,
J. Zhang⁴, W. Troy Tucker¹, Lev Ginzburg¹ and William L. Oberkampf⁵

¹Applied Biomathematics
100 North Country Road
Setauket, New York 11733
631-751-4350, fax -3435

⁴Department of Computer and Electrical Engineering
3215 Coover Hall
Iowa State University
Ames, Iowa 50014 USA
501-575-5590, fax -5339

²Department of Mathematical Science
Lewis & Clark College
0615 SW Palatine Hill Road
Portland OR 97219-7899
503-768-7560, fax -7668

⁵Validation and Uncertainty Estimation Department
MS 0828, Department 9133
Sandia National Laboratories
Albuquerque, New Mexico 87185-0828
505-844-3799, fax -4523

³Department of Ecology and Evolution
Stony Brook University
Stony Brook, New York 11794
631-632-8600, fax -7626

7
8
9
10
11
12
13
14
15

Corresponding author email: scott@ramas.com

Running head: Myths about correlation and dependence in risk analysis

Submitted to *Human and Ecological Risk Assessment*

1

2 **ABSTRACT**

3 This paper summarizes several of the most pervasive and pernicious myths about correlations and
4 dependencies that interfere with conducting good assessments. Some of these myths are obvious
5 to analysts trained in statistics and widely recognized even though they are still often perpetuated
6 for the sake of mathematical convenience. These myths include the idea that all variables in any
7 risk assessment are independent or can be treated as though they are independent. Other myths
8 are subtler or even rather esoteric and do not seem to be widely appreciated by the risk analysis
9 community. These include the ideas that Monte Carlo simulations account for dependence in a
10 comprehensive way and that varying correlation coefficients is an effective sensitivity analysis
11 when dependence between variables is unknown. Several of the myths about correlations and
12 dependencies outlined here can lead to profound errors in probabilistic risk assessments. There
13 are techniques available, however, to ensure that a proper accounting is made of intervariable
14 dependencies, even when information to specify those dependencies is lacking.

15

16

17

18

19 **KEYWORDS**

20 correlation; dependence; copula; uncorrelatedness; myths;

21

1

2 INTRODUCTION

3 Models used in probabilistic risk assessments take two kinds of inputs: (1) the marginal
4 distributions for the different variables and (2) the dependencies between these variables. The
5 second set of inputs is perhaps just as important as the first, but dependence has received
6 considerably less attention by theorists and practitioners in risk analysis. Several recent reviews
7 have considered strategies to model inter-variable dependencies in probabilistic models (Helton
8 and Davis 2003; 2002; Henderson et al. 2000; Haas 1999; Clemen and Reilly 1999; Cullen and
9 Frey 1999; Cario and Nelson 1997; Cooke 1997; Smith et al. 1992; Hutchinson and Lai 1990;
10 Morgan and Henrion 1990). Despite this literature, there are several widely held myths about
11 dependence that confuse analysts, perhaps the most pernicious of which is that it is okay to ignore
12 correlations and dependencies altogether. Even analysts who recognize the importance of
13 dependence sometimes ignore the issue because of a lack of relevant empirical data on which to
14 base a reasoned model.

15 This paper is organized as a series of debunked myths about the issue of dependence
16 among random variables in models based on mathematical functions of probability distributions.
17 It reviews the methodological dangers of assuming all variables in an assessment are independent
18 of one another and shows how different dependencies can lead to quantitatively different results.
19 It describes several strategies that could be employed to represent knowledge about how the
20 random variables are interrelated. It also includes a discussion of how the very concept of
21 independence disintegrates into distinct notions in the context of imprecise probabilities.

22 A standard approach in probability theory for modeling a joint distribution has been to
23 specify the joint distribution as a product of marginals and conditional distributions (Clemen and
24 Reilly 1999). In this way, arbitrary intervariable dependencies can be expressed in terms of
25 conditioning, at least in principle. For instance, it may be convenient to use distributions that are
26 conditional on the values sampled from other distributions. This approach has been useful in
27 hierarchical simulations (e.g., Voit et al. 1995). This strategy extends to making the parameters
28 or even the shape of a distribution depend on the value of other random variable(s). The task of
29 specifying all the necessary conditional distributions grows combinatorially with the number of
30 variables, and Clemen and Reilly (1999) suggest that this may make the approach unwieldy for
31 large assessment problems. Unless most of the underlying variables are independent or

1 conditionally independent, this strategy is information-intensive and may not often be practical
2 for risk assessments where empirical knowledge is limiting (see Myth 14).

3 The rest of this paper consists of discussions of sixteen myths that can seriously impede
4 the construction of accurate risk assessments. Additionally, we draw a few general conclusions
5 about how analysts should approach the issue of correlations and dependencies in their
6 assessments.

8 **Myth 1**

9 **All variables are mutually independent.**

10 Many variables in complex natural and engineered systems are, in fact, correlated or have some
11 nonlinear interdependence. Although most risk analysts recognize that it is improper to assume
12 that variables are independent in the face of evidence that they are not, many do so anyway as a
13 shortcut or mathematical convenience. In some cases, these counterfactual assumptions are
14 laughable, as in the case of assuming the mass of some component and its surface area are
15 independent. In some cases, assuming a perfect or opposite dependence would be a better default
16 assumption than independence. In general, it is incumbent on the analyst to model the
17 dependence if only approximately.

18 There is also impropriety in cases where independence is routinely assumed when there
19 are no observations or other evidence available about the dependence between variables one way
20 or the other. The lack of evidence about dependence does not by itself justify an assumption of
21 independence, although many analysts argue as though it does. *Fact: wishing variables were*
22 *independent so the analysis is easier doesn't make them so.* In cases when the dependence is
23 partially or completely unknown, appropriate methods to account for this epistemic uncertainty
24 should be employed.

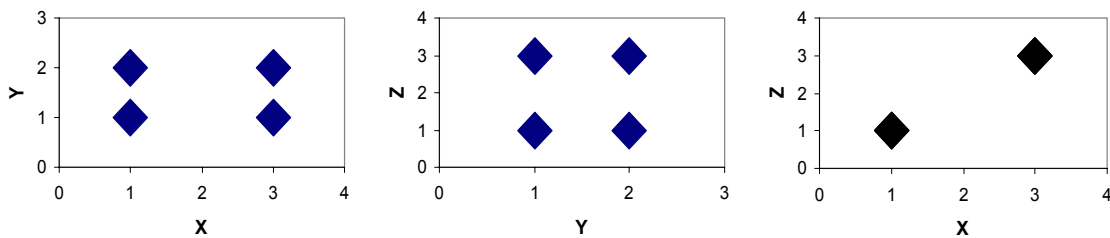
26 **Myth 2**

27 **If X and Y are independent and Y and Z are independent, then X and Z are too.**

28 Mutual independence between X and Y and between Y and Z doesn't guarantee that X and Z
29 are also independent. In other words, independence is not transitive. This fact should perhaps be
30 obvious to risk analysts. We have nevertheless observed the corresponding faulty reasoning
31 applied in actual risk assessments. *Fact: independence between X and Y and between Y and*
32 *Z implies nothing at all about the dependence between X and Z .* Consider the following very

1 simple example. Let (X, Y, Z) be a discrete distribution consisting of the four points $(1,1,1)$,
 2 $(1,2,1)$, $(3,2,3)$ and $(3,1,3)$, each with probability $\frac{1}{4}$. As depicted in Figure 1, plotting the three
 3 bivariate scattergrams (X versus Y , Y versus Z , and X versus Z) reveals that X and Y are
 4 independent, as are Y and Z , but that X and Z are (perfectly) positively dependent on each other.
 5 If the four equiprobable points of the discrete distribution are instead $(1,1,3)$, $(1,2,3)$, $(3,2,1)$ and
 6 $(3,1,1)$, then the first two graphs are unchanged, but the third graph would show an oppositely
 7 dependent relationship between X and Z .

8 It is easy to construct examples in which X is independent of Y and Y is independent of Z ,
 9 but where X and Z have any arbitrary relationship. Let F denote the distribution function for Y
 10 and let G denote the joint distribution function for X and Z (which may have any structure at all).
 11 If the trivariate joint distribution function is $H(x, y, z) = F(y) G(x, z)$, then Y is independent of
 12 both X and Z , although X and Z are themselves related in any way one might care to specify.
 13



14

15 **Figure 1. Discrete example of the non-transitivity of independence.**

16

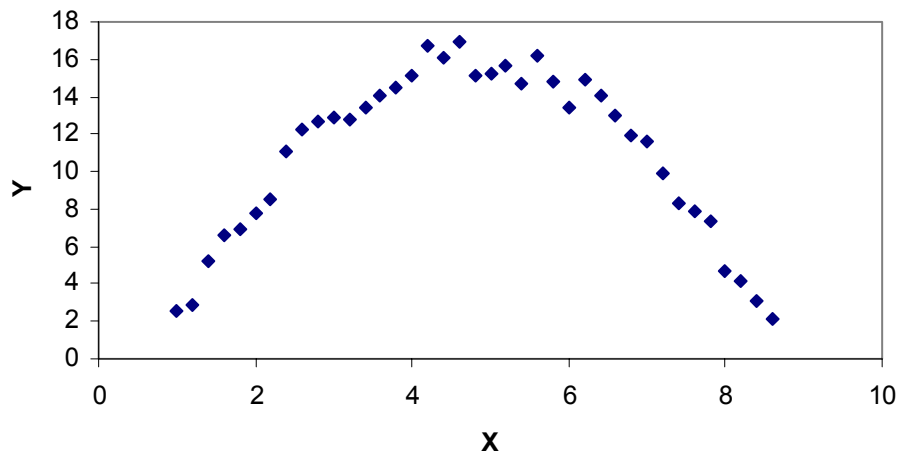
17 **Myth 3**

18 **Variables X and Y are independent if and only if they are uncorrelated.**

19 Whenever correlation is introduced in beginning statistics courses, a counterexample to this myth
 20 such as that shown in Figure 2 is immediately presented. The variables X and Y in this graph are
 21 uncorrelated, i.e., they have a Pearson correlation of zero. However, they are clearly not
 22 independent. Despite widespread attempts to disabuse students of the difference between
 23 uncorrelatedness and independence, this myth or the consequences of the myth nevertheless

1 pervade risk assessment. Uncorrelatedness does not generally* imply independence. *Fact:*
2 *independence implies that the correlation will be zero, but not vice versa.*

3 Sometimes the myth appears as “If two variables are normally distributed, then a zero
4 correlation between them implies independence.” Actually, normality of the marginal
5 distributions is not sufficient. Melnick and Tenenbein (1982) provides counterexamples (see also
6 Flury 1986; Kowalski 1973). *Fact: if variables are multivariately normal, then zero correlation*
7 *implies independence.*



8
9 **Figure 2. Variables that are uncorrelated but obviously dependent.**

10

*There are exceptions where uncorrelatedness actually does imply independence. One exception, for instance, is when X and Y both have Bernoulli distributions so that $P(X=0) = P(X=1) = 0.5$ and $P(Y=0) = P(Y=1) = 0.5$. Let $h(x,y)$ denote the joint mass function for X and Y . Let $a = h(0,0)$, $b = h(0,1)$, $c = h(1,0)$, $d = h(1,1)$, so $0 \leq a,b,c,d \leq 1$ and $a+b+c+d = 1$. Because the marginals are Bernoulli distributions, we know that $a+b = c+d = a+c = b+d = 0.5$. If X and Y are uncorrelated, then $r = E(XY) = E(X) E(Y) / \sqrt{V(X) V(Y)} = 0$, which implies $E(XY) = E(X) E(Y)$. But $E(XY) = \sum xy h(x,y) = 0 \times 0 \times a + 0 \times 1 \times b + 1 \times 0 \times c + 1 \times 1 \times d = d$. At the same time, $E(X) E(Y) = 0.5 \times 0.5 = 0.25$, so $d = 0.25$. But this means that b has to also equal 0.25 (because $b+d = 0.5$), and, in fact, the Bernoulli constraints cascade so that $a = b = c = d = 0.25$, which means that h is necessarily the independence copula. Thus, in this exceptional and highly constrained case, uncorrelatedness implies independence.

1 **Myth 4**

2 **Zero correlation between X and Y means there's no relationship between X and Y .**

3 This myth is closely related to the previous one. The phrase “no relationship” is really just
4 another way of saying that knowing the value of either variable doesn't help in any way to
5 establish the value of the other variable. Figure 2 also provides a counterexample to this myth.
6 *Fact: uncorrelatedness does not imply there is no relationship between the variables.* In Figure
7 2, X and Y are uncorrelated, but they clearly have a very strong relationship. Knowing that X is 3
8 tells us that Y is around 12. Knowing that X is 5 tells us that Y is around 15. Knowing that Y is 8
9 tells us that X is either around 2 or around 7.5. There is an immense amount of information
10 embodied in the relationship between the two variables even though they have zero correlation.
11
12

13 **Myth 5**

14 **Small correlations imply weak dependence.**

15 Figure 2 also disproves this myth, which is closely related to the previous two myths. The falsity
16 of this one is just as obvious, and yet it appears surprisingly often in multivariate data analyses
17 and risk assessments. *Fact: a weak correlation does not guarantee a weak relationship.*
18
19

20 **Myth 6**

21 **Small correlations can be “safely ignored” in risk assessments.**

22 In an important paper, Smith et al. (1992) suggested that small-magnitude correlations could be
23 “safely ignored” in risk assessments seeking estimates of means of linear arithmetic functions of
24 random variables. This is possible because *means* of sums and products are often similar to
25 means for the independent case if a simple dependence with small correlation is introduced
26 between the inputs. In the real world, however, there are three complications that prevent us from
27 ignoring dependence among variables. First, many of the functions we need to evaluate are
28 nonlinear. Second, the dependencies involved are more complicated than can be captured with
29 simple correlation coefficients (Myths 3-5<<not hyperlinked>>). Third, and probably most
30 important, risk analysts are usually more concerned about the distributions' *tails* rather than their
31 means. Tail risks can be radically influenced by dependencies even if correlation is zero (Ferson
32 et al. 2004). The Smith et al. (1992) paper has been widely overextended and abused, and risk

1 analysts should generally try to account for all dependencies that relate their input variables to
2 one another even if they might happen to yield correlations of small magnitude.

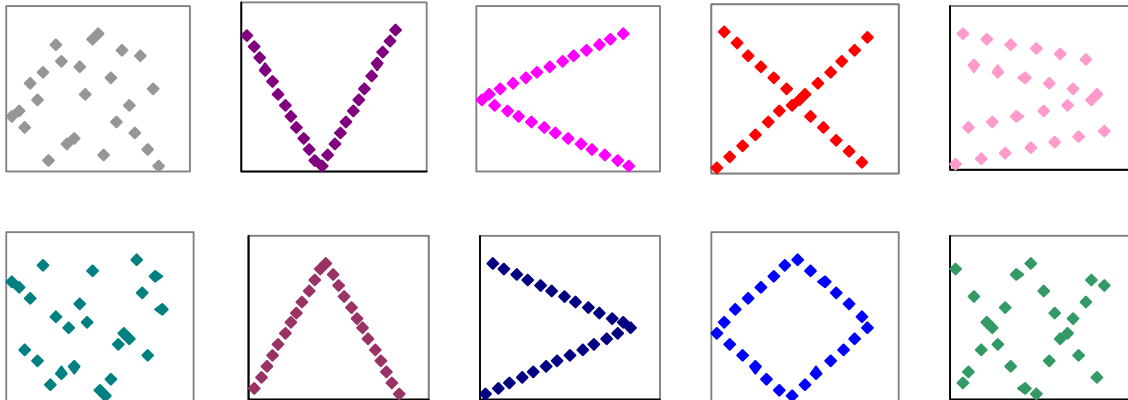
3 Let's suppose that we have a set of variables that are mutually uncorrelated, that is, their
4 pairwise Pearson correlation coefficients are all zero. It is widely understood among risk analysts
5 that uncorrelatedness does not imply independence, but the full import of this caveat is not quite
6 so widely appreciated. One might expect or hope that dependence, at least in the case when
7 correlations are all zero, would have a negligible or small effect on convolutions.* This turns out
8 to be false. Although most risk assessors recognize that uncorrelatedness between variables does
9 not formally imply their independence, many are apparently not aware of how much difference
10 dependence can make in their calculations. Many patterns of dependence produce the same
11 correlation, and, in particular, there are a *lot* of ways a joint distribution can yield uncorrelated
12 variables.

13 Consider an example in which X and Y have the same marginal distribution, which is a
14 discrete uniform on the integers from 1 to 25. Thus, the chance that X is 1 is $1/25$; the chance
15 that it's 2 is $1/25$, and so on, and the same for Y . What can be said about the sum $X+Y$ if we
16 suppose that X and Y are uncorrelated? Consider the ten dependence patterns in Figure 3. The
17 abscissa of each plot is the value of X and the ordinate is the value of Y . Because the distributions
18 are discrete, there is a mass (of size $1/25$) allotted for each of twenty-five columns in each plot.
19 Likewise, the same amount of mass is allotted for each of twenty-five rows. To make the
20 illustration easy to understand, let's further suppose that all of each row's mass is concentrated
21 into a single slug of density located at some x -value, and all of each column's mass is likewise
22 condensed at one y -value. (This is different from our previous assumption that the mass in the
23 marginal distributions were at discrete points. We're not only saying that the mass has to be at
24 the integer points, but also that there is only one y -value that has mass for each x -value.) By
25 rearranging these masses on a 25×25 grid, we can create different joint distributions between X
26 and Y . We will consider only those distributions that respect the marginal distributions by
27 constraining our arrangements so that each row has only one mass and each column has one
28 mass. Of these, we consider only those patterns that also have a correlation equal to zero (or so

*Convolution is the mathematical operation that finds the distribution of a sum of independent random variables from the distributions of the addends. The term can be generalized to refer to similar operations with differences, products, quotients, etc., and to these operations when the random variables are not independent. The distributions that result from convolutions are sometimes called "derived" distributions.

1 close to zero as to be appropriate for our example). Even with all of these constraints, there are
2 still many possible arrangements. Figure 3 depicts only a few of them.

3



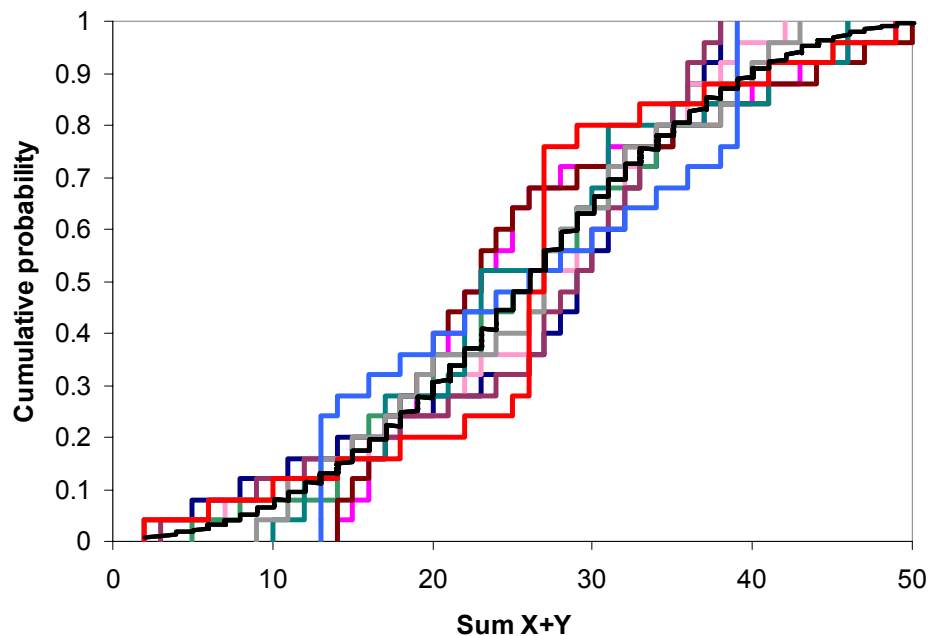
4

5 **Figure 3: Some possible patterns of dependence between uncorrelated variables.**

6

7 Now consider how these patterns of dependencies, all of which are uncorrelated,
8 influence the distribution of an arithmetic combination of X and Y . Figure 4 shows the
9 distributions of $X+Y$ associated with each of the ten patterns of dependence shown in Figure 3.
10 Also shown in this figure is the distribution under independence (it's the one going down the
11 middle with somewhat smoother tails). It should probably not be surprising that $X+Y$ depends on
12 the dependence between X and Y , but many analysts are surprised to see the magnitude of its
13 potential influence. Note, for instance, that the smallest possible value of the sum ranges between
14 2 and 14, depending on which pattern of dependency exists between the addends. This range is a
15 quarter of the entire support of the distribution! Around the value 14, the cumulative probability
16 ranges between zero and almost 30%. In other words, there might be no chance that the sum is
17 smaller than 14, or there might be a 30% chance that it's smaller than 14.

18



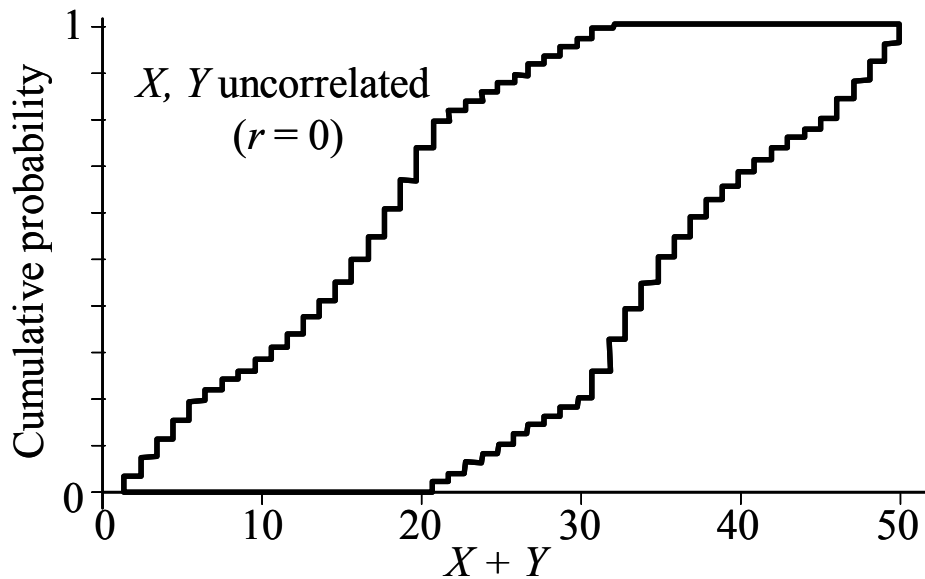
1

2 **Figure 4: Some possible distributions of a sum of uncorrelated random variables.**

3

4 In fact, the true uncertainty about the distribution of the sum is even larger than that
 5 depicted in Figure 4 which only depicts distributions for the independent case and 10 arbitrarily
 6 chosen dependent cases. There are many more patterns of dependency that would lead to
 7 uncorrelated variables. For instance, the mass need not be concentrated in unit slugs in the joint
 8 distribution. A column's mass could be distributed throughout the column without altering the
 9 discreteness of the distributions. The results depicted in Figure 4 are only a few of the infinitely
 10 many possible outcomes that are consistent with the uncorrelatedness of X and Y and their given
 11 marginal distributions. It can be shown using recently developed mathematical techniques
 12 (Makarov 1981; Frank et al. 1987; Williamson and Downs 1990; Cossette et al. 2001) that the
 13 region depicted in Figure 5 represents bounds on all distributions of the sum $X+Y$ that could arise
 14 when X and Y are uncorrelated and both distributions are uniform on the integers from 1 to 25. <<I
 15 worry that bounds on the sum could be tighter than we've shown if the inputs are discrete on the
 16 integers. Maybe the example should be transformed to the continuous case just so there's no such
 17 wrinkle possible.>> We see in Figure 5 that the minimum value of the sum can be any integer
 18 between 2 and 21, and there could be as much as a 40% chance that the sum is less than 14. All
 19 of this uncertainty surrounds the sum of X and Y , even though their marginal distributions are
 20 precisely specified and *even though the variables are exactly uncorrelated*.

1



2

3 **Figure 5: Bounds on the distribution of the sum $X+Y$ given that X and Y are**
4 **uncorrelated and identically distributed as discrete uniforms on [1,25].**

5

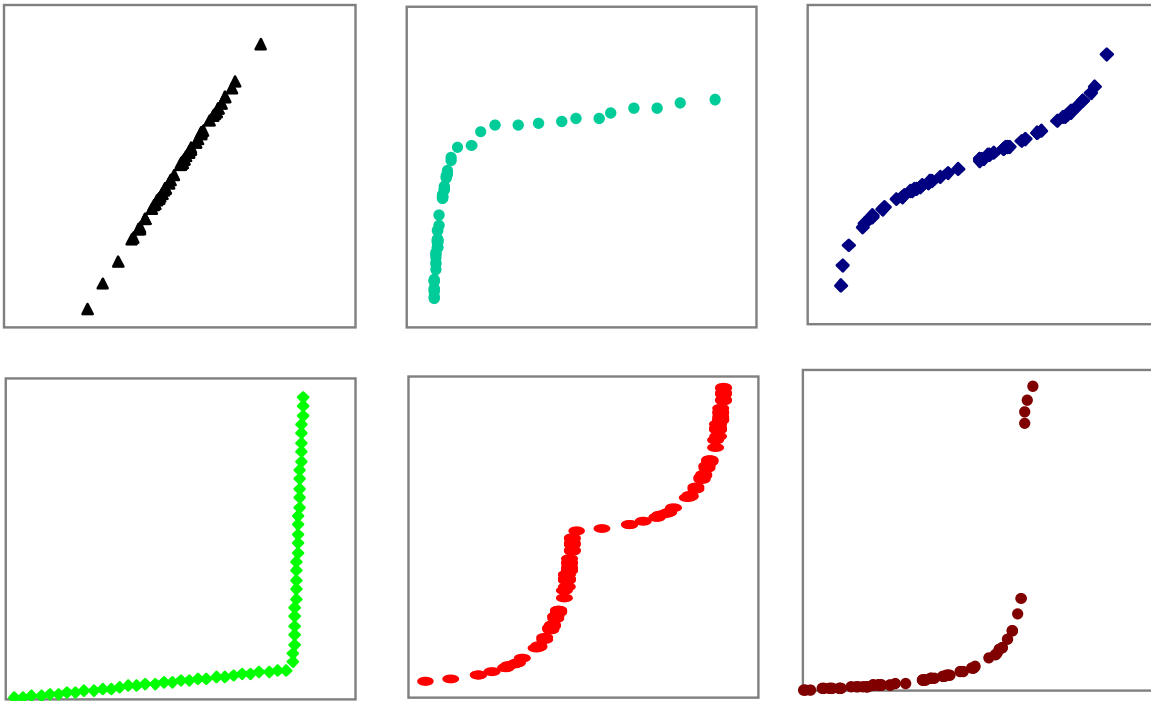
6 The implication of the remarkable breadth of uncertainty about the tail probabilities even
7 though the addends are known to be uncorrelated suggests that, in general, making unjustified
8 assumptions about independence among variables in a risk assessment can be harmful. *Fact:*
9 *even negligible correlations can greatly influence the results of risk calculations.* This is why
10 Myth 1 [is important](#). Many risk analysts reflexively assume independence
11 among all events or random variables even when they have no particular justification for doing so
12 other than mathematical convenience. It is improper, however, to assume independence among
13 variables in an analysis unless there is reliable evidence or a compelling argument that this is a
14 reasonable assumption. If a dependency is neglected, the answer obtained by an analysis
15 assuming independence will generally be wrong. Under certain conditions, the central tendency
16 of output distributions could be approximately correct (Smith et al. 1992). However, the
17 estimated dispersion and especially the tail probabilities can be highly inaccurate (Bukowski et al.
18 1995; Ferson and Burgman 1995; Ferson 1994; Ferson et al. 2004). In some cases, the
19 dispersion will be larger than it should be. In some cases, it will be smaller. In the latter, the
20 probabilities of extreme events will likely be underestimated. These extreme events are often the
21 primary focus of the risk assessment. They may represent very large stresses or threatening
22 conditions that correspond to system failures and structural collapses that the risk analysis was

1 intended to assess. In such cases, it is therefore crucial that these tail probabilities be accurately
2 characterized and, in no circumstance, underestimated. Assuming independence without proper
3 justification amounts to wishful thinking and is therefore detrimental to the purposes of risk
4 assessment to be a dispassionate and reasoned accounting of the possible adverse consequences
5 and their probabilities. Independence assumptions must also be used in a consistent manner. For
6 instance, if the random variable exposure is a function of concentration and intake, then it cannot
7 be independent of either of these variables. Such inconsistent assumptions sometimes arise in
8 backcalculation analyses seeking to establish cleanup targets. They generally lead to nonsensical
9 results.

11 **Myth 7**

12 **Different correlation coefficients are similar.**

13 Some risk analysts suggest that it doesn't make much difference which measure of correlation is
14 employed and that the various measures are pretty much interchangeable. This view is false,
15 however, as even cursory inspection of examples will easily reveal. There are many different
16 measures of correlation that are in common use and many more that have been proposed. The
17 most commonly used measures are Pearson's product-moment correlation and Spearman's (1904)
18 rank correlation, but there are a host of other measures that also arise in various engineering
19 contexts, including Kendall's rank correlation, concordance of various kinds (e.g., Hoeffding
20 1947; Lehmann 1966; Scarsini 1984), Blomqvist's (1950) coefficient, Gini's coefficient (Nelsen
21 1999), etc. Hutchinson and Lai (1990) review many of these. The choice of the measure can
22 strongly influence the numerical characterization of a scattergram. Figure 6 shows a variety of
23 bivariate relationships as scattergrams. Note that the units of the abscissa and ordinate are not
24 shown because they are irrelevant and do not affect the magnitudes of the correlations. Each of
25 the six scattergrams displayed has the same Spearman rank correlation, which is one,
26 corresponding to perfect dependence or comonotonicity. But the scattergrams have widely
27 different Pearson correlation coefficients. For instance, the Pearson correlation for the
28 scattergram in the upper, left-hand graph is one, but the Pearson correlation for the scattergram
29 below it in the lower, left-hand graph is about 0.6. *Fact: the various measures of correlation are*
30 *sensitive to different features of the scattergram.*



1

2

Figure 6. Different bivariate relationships with the same Spearman rank correlation (unity) but widely different Pearson correlation coefficients.

3

4

5 **Myth 8**

6 **A correlation coefficient specifies the dependence between two random variables.**

7 *In fact, it takes an entire dependence function or “copula” to fully specify the dependence*
 8 *between two random variables* (Genest and MacKay 1986; Hutchinson and Lai 1990; Schweizer
 9 1991; Dall’Aglio et al. 1991; Nelsen 1991; 1995; 1999; Haas 1999; Clemen and Reilly 1999; Li
 10 2000). A correlation coefficient is often a very poor summary of the dependence; it generally
 11 does *not* specify or determine the dependence. Instead, it determines only a class of such
 12 dependencies. In other words, many dependence functions have the same correlation. See Haas
 13 (1999), Genest and MacKay (1986) or Li (2000) for an introduction to copulas; see Hutchinson
 14 and Lai (1990), Nelsen (1999) or Dall’Aglio et al. (1991) for a monographic treatment.

15 Copulas are simply the dependence functions that knit together marginal distributions to
 16 form their joint distribution. The copula between X and Y is just the joint distribution function
 17 between the uniformly distributed variables $F_X(X)$ and $F_Y(Y)$, where F_X is the distribution function
 18 for X and F_Y is the distribution function for Y . Mathematically, a copula is a function $C: [0,1] \times$
 19 $[0,1] \rightarrow [0,1]$ such that $C(a, 0) = C(0, a) = 0$ and $C(a, 1) = C(1, a) = a$ for all a in $[0,1]$, and $0 \leq$

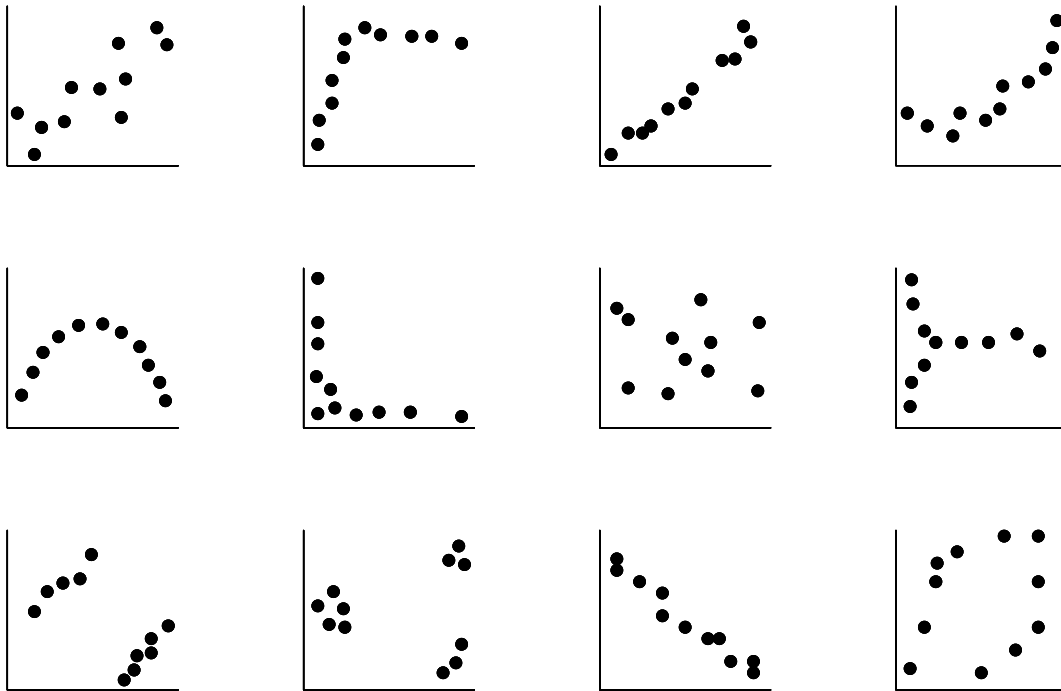
1 $C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1)$ for all a, a_1, a_2, b_1, b_2 in $[0, 1]$, when $a_1 \leq a_2$ and $b_1 \leq b_2$.
2 Copulas have relatively simple structures (at least compared to joint distributions in general) and
3 have many useful properties. For instance, they greatly simplify the generation of correlated
4 random numbers (see, e.g., Nelsen 1986; 1987; Clemen and Reilly 1999). Every dependence
5 between or among random variables, even including functional relationships, is expressed by
6 some copula. Sklar's (1959) theorem tells us how to compute the joint distribution function
7 $H(x, y)$ from specified marginal distributions and a dependence function represented by a copula.
8 For any (univariate) distribution functions F_X and F_Y and any copula C , the function

$$H(x, y) = C(F_X(x), F_Y(y))$$

11
12 is a two-dimensional distribution function having marginals F_X and F_Y . This is a decomposition
13 of a joint distribution into its marginals and the copula that knits them together (Sklar 1959;
14 Schweizer 1991; Nelsen 1999). Sklar's theorem generalizes to dimensions higher than two
15 (Nelsen 1999; Cossette et al. 2001).

16 It should be obvious that any function like a copula cannot be completely characterized
17 by a single dimension such as correlation coefficient. Iman and Davenport (1980; 1982a)
18 illustrate dozens of bivariate patterns of great variety, but all of the pictures they show are
19 extremely simple compared to the host of patterns that are possible between two variables. In
20 fact, there are many, many ways for dependence between variables to be expressed. There can be
21 all manner of nonlinear relationships, clusters, subgroupings, impossible regions, and other
22 complexities. Figure 7 can only hint at the possible variety. Hart et al. (2004) argue that some of
23 these nonlinear patterns actually arise in risk assessment problems. For instance, two swaths of
24 points such as those shown in the lower, leftmost scattergram might arise from plotting both
25 males and females on a graph of physiological variables. Loops similar to that shown in the
26 lower, rightmost scattergram arise so commonly in data that they have been given several names,
27 including "Kendall's arch" in ecology, the "circumplex" in psychology, and the "horseshoe" in
28 archeology (Wartenberg et al. 1987). Many of the bivariate patterns illustrated in Figure 7 are
29 only very poorly captured by a scalar correlation coefficient. A copula, on the other hand, can
30 completely capture the dependence between any variables.

31



1

2 **Figure 7. A handful of bivariate scattergram patterns, only some of which can be**
 3 **well characterized by any one-dimensional correlation coefficient.**

4

5 Despite all appearances, the previous Figure 6 is actually not an illustration of the fact
 6 that many dependence functions have the same correlation. All of the scattergrams in this figure
 7 have exactly the same dependence function. Their dependence is perfect in that each random
 8 variable is a non-decreasing function* of the other. This dependence is expressed by the
 9 particular copula $M(u,v) = \min(u,v)$. Some authors call this relationship comonotonicity (e.g.,
 10 Müller 1997; Goovaerts et al. 2000; Kaas et al. 2000). The differences among the scattergrams in
 11 Figure 6 are due entirely to the differences in the marginal distributions for the abscissa and
 12 ordinate variables.

13 In the assertion above that the correlation does not generally specify the dependence, the
 14 adverb “generally” was necessary because there some exceptions when the correlation does
 15 completely determine the dependence. One exception is the case discussed above of the
 16 Bernoulli marginals where uncorrelatedness implies independence. Another exception is when
 17 the correlation is extreme, that is, when the dependence is perfect or opposite. In this case, the

*Assuming variables are perfectly dependent (comonotonic) is different from assuming that either is completely dependent on the other, which is a more general situation.

1 Spearman rank correlation and Kendall correlation are either +1 or -1 (and the Pearson
2 correlation is as large or as small as it can get given the marginal distributions). When one of
3 these correlation measures is ± 1 , the dependence function is fully determined. Interestingly, as
4 the correlation gets closer to zero, the family of dependence functions having that correlation gets
5 larger and larger. (This fact tends to explain why Myth 6 is not true.) See Figure 3 for examples
6 of scattergrams corresponding to fixed marginal distributions with different dependence functions
7 that all have the same Pearson correlation.

8

9 **Myth 9**

10 **Correlation varies between -1 and +1.**

11 By convention, most measures of correlation are scaled so that they range in the interval $[-1, +1]$.
12 Some measures, such as Spearman correlation, can always range over this entire interval. But not
13 all correlation coefficients can vary across this range for arbitrary marginal distributions. The
14 Pearson correlation, in particular, often cannot achieve either -1 or +1. For instance, if X is
15 uniformly distributed over the unit interval $[0,1]$ and Y is a lognormal distribution with underlying
16 $\mu = 0$ and $\sigma = 1$, then the correlation between X and Y cannot be any larger than about 0.7.
17 Depending on the marginal distributions involved, the largest possible Pearson correlation could
18 in fact be arbitrarily close to zero ($\ll \diamond \gg$). *Fact: the Pearson correlation coefficient ranges*
19 *within $[-1, +1]$, but it may not reach all possible values in the interval for some marginal*
20 *distributions.*

21

22 **Myth 10**

23 **Any correlation can be specified between inputs.**

24 There are mathematical constraints associated with correlations that forbid certain combinations.
25 For instance, one variable cannot be strongly positively correlated with each of two variables that
26 are themselves strongly negatively correlated. Such constraints can be summarized by saying the
27 matrix of correlations must always be a positive semi-definite matrix (Eves 1966). Checking for
28 positive semi-definiteness requires a special algorithm (Iman and Davenport 1982b; Ferson
29 1996a). If correlations from different studies are mixed into a single analysis, or if correlations
30 are based on hypothetical values or best professional judgment, infeasible configurations may be

1 specified. *Fact: the pairwise correlations for a set of variables must satisfy certain feasibility*
2 *constraints, so not all sets of correlations that one might specify are possible.*

3 If the matrix is positive semi-definite, then it is a possible correlation matrix. If it is not
4 positive semi-definite, then it cannot be a correlation matrix in the first place and certainly should
5 not be used in modeling dependencies in a risk analysis. A model that uses an infeasible
6 correlation matrix will produce gibberish. Many specially developed computer codes and even
7 some commercially available software packages for Monte Carlo simulation do not check that the
8 input correlation matrix satisfies the positive semi-definiteness condition, e.g., early versions of
9 @Risk (Palisade Corporation 1996). Consequently, they will produce nonsensical results
10 whenever users specify an infeasible set of correlations. It is therefore important for analysts
11 always to check that the input corresponds to a feasible correlation matrix.

12 It is possible and potentially useful to employ the positive semi-definiteness of
13 correlation matrices to tighten interval estimates of correlation. For instance, knowing some of
14 the correlations between four variables W , X , Y and Z can inform us about the others. Suppose
15 the matrix

16

	W	X	Y	Z
W	1	0	0.7	0.8
X	0	1	0	?
Y	0.7	0	1	?
Z	0.8	?	?	1

21

22
23 characterizes the pairwise correlations, where the question marks mean we don't know the
24 correlation for those variable pairs. An obvious trial-and-error algorithm that tests possible
25 values for the unknown correlations tells us that the correlation between X and Z must be within
26 the range $[0, 0.6]$, and the correlation between Y and Z must be in the range $[0.13, 0.99]$. Any
27 values outside these ranges would violate positive semi-definiteness of the correlation matrix.
28 Thus, the knowledge about the magnitudes of some of the correlations allows us to impute
29 something about the correlations of the others.

30

1 **Myth 11**

2 **Perfect dependencies between X and Y and between X and Z imply perfect**
3 **dependence between Y and Z .**

4 Extending the ideas discussed above about constraints on the correlation matrix, one might have
5 expected that if a variable X is maximally correlated to variable Y , and variable Y is maximally
6 correlated to variable Z , then we might be able to conclude that X and Z are also maximally
7 correlated. Expressed in other terms, comonotonicity between both X and Y and between Y and Z
8 would seem to imply there should likewise be comonotonicity between X and Z . Furthermore, one
9 might expect that if X and Y are maximally correlated (comonotonic) and Y and Z are minimally
10 correlated (sometimes called “countermonotonic”) then X and Z should be minimally correlated
11 (countermonotonic) too.

12 Unfortunately, this strategy of using available information about the relationships among
13 some variables to inform us about the relationships among others does not extend to feasibility
14 constraints on the qualitative (sign) information about dependencies, which is especially weak.
15 Even information about how some variables are perfectly or oppositely dependent does not induce
16 constraints that can be used to make inferences about unknown dependencies. Indeed, seemingly
17 self-evident inferences involving extremal dependencies are demonstrably false. For instance,
18 suppose A and B are oppositely dependent and that A and C are oppositely dependent. Thinking
19 something like “the enemy of my enemy is my friend”, one might expect that it would be possible
20 to infer from this that B and C are perfect dependent. However, this is not a correct inference.
21 Although one can infer that B and C could not be oppositely dependent, they may be independent.
22 Here is a simple example. Consider discrete distributions such that there are four possible
23 configurations as given in the following table.

24

	A	B	C
25			
26	1	3	3
27	2	1	3
28	2	3	1
29	3	1	1
30			

31 It is easy to see by plotting these three variables against each other in various combinations, that
32 A and B are oppositely dependent on one another, as are A and C . (Their Pearson correlation is
33 -0.707 , but their Spearman correlation is -1 .) Nevertheless, B and C are independent. Likewise,
34 it is very easy to construct examples in which a variable X is perfectly associated with both Y and

1 Z, and yet the variables Y and Z themselves are independent. Thus, one cannot use information
2 that some variables are maximally dependent to infer very much about other variables.

3 Let $//$ denote perfect dependence, i.e., maximal correlation and comonotonicity, and let $\backslash\backslash$
4 denote opposite dependence, i.e., minimal correlation and countermonotonicity. Below are facts
5 that correct some of the mistaken ideas:

6
7 Fact: $X//Y$, and $Y//Z$ do not imply $X//Z$.

8 Fact: $X//Y$, and $Y\backslash\backslash Z$ do not imply $X\backslash\backslash Z$.

9 Fact: $X\backslash\backslash Y$, and $Y\backslash\backslash Z$ do not imply $X//Z$.

10
11 Perhaps even more surprising is that $X//Y$ and $Y//Z$ together don't even imply that X and Z can't
12 be independent. A counterexample is easy to construct. Let (X, Y, Z) be discrete, taking on of the
13 four values $(1,1,1)$, $(1,2,3)$, $(3,2,1)$, and $(3,3,3)$, each with probability $\frac{1}{4}$. Sketching the three
14 bivariate plots reveals that $X//Y$ and $Y//Z$, but X and Z are independent. It is possible to
15 conclude from perfect dependence between X and Y and between Y and Z that X and Z cannot be
16 oppositely dependent, but that is a fantastically weaker conclusion that will rarely matter in a
17 practical risk assessment.

18 If Y and Z are independent, then $f(Y)$ and $g(Z)$ are also independent, where f and g are
19 arbitrary measurable functions (Roussas 1997, page 166). One might expect this fact could be
20 extended to comonotonic or countermonotonic variables, but this is not the case. Let \perp denote
21 independence.

22
23 Fact: $X//Y$, and $Y\perp Z$ do not imply $X\perp Z$.

24 Fact: $X\backslash\backslash Y$, and $Y\perp Z$ do not imply $X\perp Z$.

25
26 The combination of perfect dependence with independence is subtle, and the mistakes that
27 analysts make are understandable. In fact, however, assuming perfect or opposite dependence
28 between X and Y and independence between Y and Z doesn't allow any conclusion at all about the
29 dependence between X and Z . Any relationship between them is possible. One example would
30 be where (X, Y, Z) take on the four values $(1,1,3)$, $(2,1,1)$, $(2,3,3)$ and $(3,3,1)$, each with
31 probability $\frac{1}{4}$. Bivariate sketches show that $X//Y$ and $Y\perp Z$, but $X\backslash\backslash Z$. If the equiprobable
32 points were instead $(1,1,1)$, $(2,1,3)$, $(2,3,1)$ and $(3,3,3)$, then still $X//Y$ and $Y\perp Z$, but now $X//Z$.

33 This flexibility about dependencies might be surprising because it seems to contradict the
34 strictures on correlations mentioned in the discussion of the previous myth. *In fact, the constraint*
35 *of positive semi-definiteness that correlations must observe does not generalize to the case of*

1 *dependencies, even in the extreme cases where correlations are minimal or maximal.* This is
2 another example of the richness of dependence in general as compared to the much narrower
3 measures of correlation.

4 5 **Myth 12**

6 **Monte Carlo simulations can account for dependencies between variables.**

7 Cullen and Frey (1999, page 202) complain that critics of Monte Carlo simulation unfairly accuse
8 it of “ignoring correlations”. They point out that restricted pairing methods developed by Iman
9 and his colleagues allow analysts to construct deviates in Monte Carlo simulations that have a
10 prescribed correlation (Iman and Conover 1982; Iman and Davenport 1982a; 1982b; Iman and
11 Helton 1985; Iman and Shortencarier 1984; Helton 1993; Helton and Davis 2002; 2003).

12 However, what Cullen and Frey don’t mention is that these algorithms pick a *particular*
13 dependency function with the prescribed correlation, and that this is only one of infinitely many
14 possible dependencies having this correlation. *Fact: Monte Carlo methods can simulate*
15 *correlations, but they do so by making unstated assumptions about the nature of the copula*
16 *representing the dependence function.* Monte Carlo methods cannot truly account for
17 correlations in the sense of computing how low or high risks might be without making such
18 assumptions. As discussed in consideration of Myth 6 the effect on numerical results of these
19 different dependence functions can be substantial, even though they may all have the same
20 correlation coefficient.

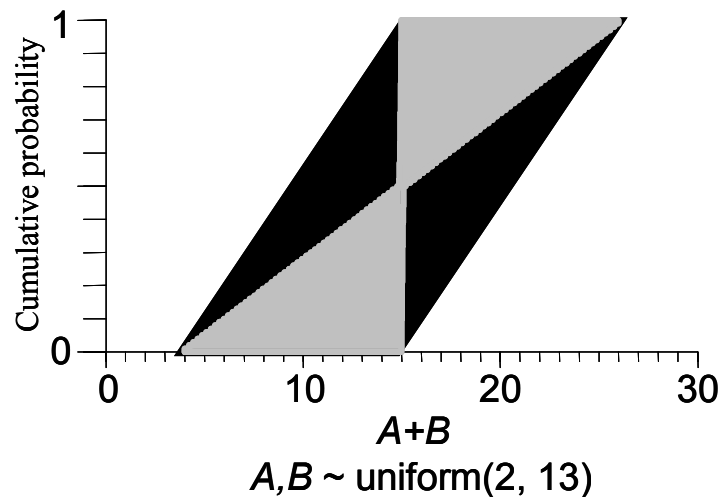
21 The origin of Myth 7 discussed above seems to be due to the mistaken impression that
22 Monte Carlo simulation fully or reasonably accounts for correlations. Given a magnitude of the
23 correlation, one observes scattergrams from Monte Carlo simulations that are fairly similar to one
24 another whichever measure of correlation is used. For instance, when specifying a correlation of
25 0.5, it makes little difference whether the Spearman or the Kendall correlation is used as the
26 resulting bivariate scattergrams are visually indistinguishable (Nelsen 1986; 1987). This
27 similarity is, of course, mostly a consequence of the very myopic selection of the dependence
28 function used in Monte Carlo simulations to generate correlated deviates. There are generally
29 other possible choices for the dependence function that have the same correlation and yet produce
30 substantially different scattergrams and result in considerably different answers.

31

1 **Myth 13**

2 **Varying correlation coefficients constitutes a sensitivity analysis for uncertainty**
3 **about dependence.**

4 *Fact: varying correlations is insufficient to explore the possible nonlinear dependencies between*
5 *variables.* For this reason, a sensitivity analysis based on varying correlation gives an incomplete
6 picture of uncertainty that is far too tight, even if we vary correlation between -1 and $+1$. As an
7 example, consider the problem of estimating the distribution of the sum $A+B$, where A and B are
8 both uniform random numbers over the interval $[2,13]$ but the dependence between A and B is
9 unknown. The range of distributions that would be seen in Monte Carlo simulations by varying
10 the correlation between A and B over its possible range of $[-1, +1]$ is shown in Figure 8 as a gray
11 slanted hourglass. This hourglass can be computed by using any Monte Carlo simulation package
12 that handles correlations such as Crystal Ball (Decisioneering 1996; Burmaster and Udell 1990;
13 Metzger et al. 1998); @Risk (Palisade Corporation 1996; Salmento et al. 1989; Barton 1989;
14 Metzger et al. 1998), or Unicorn (Cooke 2002), simply by varying the correlation between the
15 extreme possible values. This hourglass can be compared with the best possible limits on these
16 distributions computed without any assumptions about dependence. These limits form a black
17 parallelogram underneath the hourglass. The limits can be computed using the methods of Frank
18 et al. (1987; Williamson and Downs 1990; Berleant and Goodman Strauss 1998; see also Myth 15
19 below). The black parallelogram is substantially larger than the gray hourglass, and, although the
20 bulk of the difference is about the central parts of the distribution rather than the extreme tails, the
21 potential tails risks are underestimated by the Monte Carlo sensitivity analysis strategy.
22



23

1 **Figure 8. The range of distributions of sums $A+B$ obtainable from Monte Carlo**
2 **simulations varying the A,B correlation between -1 and $+1$ (gray**
3 **hourglass) and best possible limits on these distributions making no**
4 **assumptions about dependence (black parallelogram).**

5
6 **Myth 14**

7 **A model should be expressed in terms of *independent* variables only.**

8 Some methodologists (e.g., Morgan and Henrion 1990) argue that it would be best for an
9 analyst to reduce any risk assessment problem involving dependent variables into one involving
10 only independent variables by conditioning distributions as mentioned in the introduction or
11 functionally modeling the physical or biological relationships among the variables. This strategy
12 does not try to characterize the dependencies statistically, but rather sidesteps the problem
13 altogether. In the case of a risk expression involving correlated variables X and Y , this strategy
14 would replace the Y with some function of X based on the physical relationship that produced the
15 dependency between the variables in the first place. If this relationship is completely specified,
16 the value of Y can be precisely determined solely by the value of X . Of course, cases of such
17 complete predictability are very rare in science and engineering, and generally the function will
18 involve a random error term that represents the residual uncertainty about Y after accounting for
19 X . By construction, however, this error term can be made independent of X , and therefore the
20 problem with two correlated variables has been changed into a different problem with two, or
21 possibly more, independent variables.

22 Although this approach can require considerably more scientific understanding about the
23 modeled system than is commonly available in risk assessments, some analysts feel this strategy
24 is the best way to treat dependencies. For instance, the developers of the probabilistic modeling
25 software package Analytica* suggest that any dependencies present should be accounted for and
26 modeled explicitly (Morgan and Henrion 1998). In fact, their package does not even support
27 user-defined correlations, so it forces users to untangle any dependencies before they can begin
28 an analysis.

29 This purist approach does not always provide a workable strategy however. For example,
30 suppose an analyst has been charged with conducting a risk assessment for vegetation wildfire in

*Analytica is the successor to the Demos software (Morgan and Henrion 1990).

1 the Everglades that might be sparked by a malfunction and explosion of solid-propellant boosters
2 used at Cape Canaveral. Such an assessment would likely be very complex and might involve
3 considerations about current weather patterns such as a wind rose, humidity distributions, recent
4 weather's impact on the vegetation's fire risk, and a host of sundry design and mission
5 parameters. The model of the explosion's effects on the ground vegetation might require
6 probability distributions for the mass and surface area for fragments of the propellant and the
7 housing vehicle. These variables are clearly unlikely to be stochastically independent of one
8 another. A functional modeling approach to accounting for their dependence would be to develop
9 a submodel about how the fragments were produced in the explosion process itself. Obviously,
10 this could significantly enlarge the modeling effort.

11 Even if the analyst were game to undertake the challenge of modeling the generation of
12 explosion fragments, there could be other pesky correlations and dependencies among the
13 weather parameters. For instance, wind speed and humidity may not be independent
14 meteorological variables in south Florida. Vegetation fire risk tends to vary over the course of a
15 year. Therefore the timing of launches may tend to covary with fire risk on the ground. To
16 explicate all of these dependencies by functional modeling, the analyst would need to employ (or
17 become) a meteorologist. At some point, the analytical demands of the functional modeling
18 approach will likely become prohibitive. Besides the obvious disadvantage owing to the extra
19 modeling effort that may be required by the use of functional modeling to account for
20 dependencies in a risk assessment, there is one further caveat: it is not generally sufficient to
21 transform the model into uncorrelated variables (Myth 6); they must be statistically *independent*
22 variables.

23 A similar and related strategy is to use stratification to obviate the need to model
24 dependencies. Some risk analysts find it useful to stratify the assessment by creating relatively
25 homogeneous subgroups that have similar characteristics to reduce dependencies among variables
26 (Frey 1992; Cullen and Frey 1999). For these cases, one isolates the covariance into the
27 difference between the groups. Within groups, the assumption of independence is more
28 reasonable and workable. For instance, if some components were manufactured at Oak Ridge,
29 Tennessee, and some were manufactured in Paducah, Kentucky, it may be reasonable to treat
30 these two subgroups in completely separate analyses rather than trying to pool them together into
31 a heterogeneous population of components manufactured at two facilities. Such stratification by
32 age group or gender is often employed in human health assessments in part to avoid having to
33 specify and model correlations. The separate treatment of different receptor wildlife species can
34 also be viewed as an example of this strategy. The cost of this strategy is that the analysis

1 becomes more complex and cumbersome because it must be repeated for each new group in the
2 stratification.

3 Strictly speaking, the idea that a model should be expressed in terms of independent
4 variables only might better be labeled an “opinion” than a “myth”, but the idea is so clearly
5 unworkable in general that it seems fair to list it here with other ideas that are impediments to
6 conducting good risk assessments. Modeling all the underlying sources of the dependencies will
7 quickly become unwieldy and may be recursively complex. *Fact: a statistical approach is*
8 *needed to handle dependencies in most risk assessment models.*

9

10 **Myth 15**

11 **You have to know the dependence to model it.**

12 Recent algorithmic advances (Frank et al. 1987; Williamson and Downs 1990; Berleant and
13 Goodman-Strauss 1998; Cossette et al. 2001; Ferson et al. 2004; Berleant and Zhang 2004) allow
14 the calculation of bounds on risks (1) under only partially specified dependence functions, or
15 even (2) *without any assumption whatever about dependence*. Even if there is no information at
16 all available about the dependence function relating variable X and Y for which we know the
17 respective marginal distributions F and G , it is still possible to compute upper and lower bounds
18 on the distribution for $Z = X+Y$ as

19

$$20 \left[\sup_{z=x+y} \max(F(x) + G(y) - 1, 0), \inf_{z=x+y} \min(F(x) + G(y), 1) \right],$$

21

22 where sup and inf denote the supremum and infimum respectively. These limits are bounds on
23 the distribution function of the sum for every possible value z it might take. The limits are based
24 on the classical Fréchet-Hoeffding limits for the dependence (copula) function. This formula was
25 used, for instance, to compute the black parallelogram depicted in Figure 8. There are similar
26 formulas for the distribution of differences, products, quotients, etc.

27 When there is partial information about the dependence function, such as that the
28 relationship between X and Y is certainly positive (positive-quadrant dependent, Nelsen 1999;
29 Lehman 1966), then bounds on the distribution for Z can be computed with a formula like

30

$$31 \left[\sup_{z=x+y} (F(x)G(y)), \inf_{z=x+y} (1 - (1 - F(x))(1 - G(y))) \right].$$

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

There are similar formulas for differences, etc., and for dependence functions that are surely negative. See Cossette et al. (2001) or Ferson et al. (2004) for a fuller discussion of these formulas and general strategies to compute bounds on derived distribution functions when the dependence between their arguments is imprecisely known. Berleant and Zhang (2004) describe an alternative linear programming solution to this problem.

Myth 16

The notion of independence generalizes to imprecise probabilities.

In probability theory, there are several ways to define the concept of independence between events and between random variables. For events A and B characterized by real-valued probabilities $P(A)$ and $P(B)$, independence between A and B is implied by any of the following conditions:

- i) $P(A \& B) = P(A) P(B)$,
- ii) $P(A \vee B) = P(A) + P(B) - P(A) P(B)$,
- iii) $P(A | B) = P(A)$ if $0 < P(B)$,
- iv) $P(B | A) = P(B)$ if $0 < P(A)$,

where P denotes the probability of an event. It is an elementary exercise in mathematical probability to prove that each of these four conditions implies the others (Mood et al. 1974, page 40). The case of random variables similarly has several possible definitions for independence. For random variables X and Y characterized by the joint distribution H with marginals F and G such that $P(X \leq x) = F(x)$, $P(Y \leq y) = G(y)$ and $P(X \leq x, Y \leq y) = H(x, y)$, then independence between X and Y implies, and is implied by, each of the following conditions:

- i) $H(x,y) = F(x) G(y)$, for all values x and y ,
- ii) $P(X \in I, Y \in J) = P(X \in I) P(Y \in J)$, for any subsets I, J of the real line,
- iii) $h(x,y) = f(x) g(y)$, for all values x and y ,
- iv) $P(x \leq X | Y = y) = P(x \leq X)$ and $P(y \leq Y | X = x) = P(y \leq Y)$, for all x and y ,
- v) $E(w(X)z(Y)) = E(w(X)) E(z(Y))$, for arbitrary functions w and z , and
- vi) $\varphi_{X,Y}(t,s) = \varphi_X(t) \varphi_Y(s)$.

where P is probability, f , g and h are density analogs of F , G and H respectively, E is expectation, and φ denotes the Fourier transform (characteristic function). As was true for events, when probabilities are precise these various definitions of independence between random numbers are all *equivalent*. Each definition implies all the others. Therefore, there's a single concept of

1 independence, expressed by the product copula Π , that simultaneously embodies all of these
2 possible definitions.

3 There is a decidedly different story in the context of imprecise probabilities (Walley
4 1991), robust Bayesian analysis (Insua and Ruggeri 2000; Berger 1985), or Dempster-Shafer
5 evidence theory (Shafer 1976; Caselton and Luo 1992; Tonon et al. 2000), where an input
6 variable might be modeled by a *class* of probability distribution functions rather than a single,
7 precisely specified probability distribution. When probabilities are imprecise, the special case of
8 independence, which is unique in probability theory, disintegrates into several different cases.
9 Couso et al. (2000) pointed out that, for imprecise probabilities (which includes both Dempster-
10 Shafer structures and probability boxes as special cases), the various possible definitions of
11 independence are no longer equivalent to each other. The different definitions induce distinct
12 concepts of independence for imprecise probabilities. Couso et al. (2000) defined a half a dozen
13 concepts that might be called independence, including epistemic independence and strong
14 independence.

- 15 • *Epistemic independence* arises when an analyst's uncertainty about either of two
16 outcomes of a random experiment does not change when some information about the
17 outcome of one of them becomes known. Random variables X and Y are epistemically
18 independent if the conditional probability of each given the other is equal to its
19 unconditional probability, so that $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. In the context of
20 imprecise probabilities, epistemic independence is defined in terms of lower bounds on
21 expectations such that $\underline{E}(f(X)|Y) = \underline{E}(f(X))$ and $\underline{E}(f(Y)|X) = \underline{E}(f(Y))$ for all functions f where
22 $\underline{E}(Z)$ denotes the infimum of all expectations of Z over all possible probability
23 distributions that could characterize Z .
- 24 • *Strong independence* is the complete absence of any relationship between random
25 variables. Variables X and Y are strongly independent if the set of possible joint
26 distributions is the largest set such that each joint distribution $H(x, y) = F(x) G(y)$, where
27 F is one of the possible distribution functions characterizing X and G is one of the
28 possible distribution functions characterizing Y . Variables X and Y should be
29 characterized as strongly independent if (i) X and Y result from random experiments, each
30 governed by a unique albeit possibly unknown probability distribution, (ii) the random
31 experiments are stochastically independent (in the traditional sense), and (iii) there is no
32 known relationship between the variables that would preclude some possible
33 combinations of the possible marginal distributions.

1 Couso et al. (1999) gave examples of how these different definitions could alter the numerical
2 calculations for a risk calculation based on imprecise probabilities. Fetz (2001; Fetz and
3 Oberguggenberger 2004) illustrated the consequences of the various independence definitions in
4 a probabilistic assessment for an engineering system. (In the light of these differences, it is clear
5 that Myth 3 and Myth 4 are actually distinct ideas, even though they seem very similar.) Cozman
6 and Walley (2001) explored the properties of epistemic irrelevance and epistemic independence.
7 *Fact: there are several distinct concepts that could rightly be called independence in the context*
8 *of imprecise probabilities (Couso et al 1999). These concepts are not equivalent to one another*
9 *and they can lead to numerical differences in convolution results.*

10 **CONCLUSIONS**

11 One common strategy for dealing with dependencies among variables is to simply ignore them
12 and assume all variables are independent of one another. It is still common practice among risk
13 analysts in many quarters to assume independence among variables even when this assumption is
14 known to be false. The reasons for this are many, ranging from mathematical convenience and
15 the laziness of the analyst, to the preliminary nature of the analysis, to the inaccessibility of ready
16 and workable alternative strategies and misconceptions about how important dependence can be
17 in risk assessments. Despite its commonness, this strategy is clearly indefensible, and it is
18 potentially dangerous if it yields underestimates of risks of extreme adverse events.

19 The discussion of the various myths about correlation and dependence in this paper has
20 indicated that there are several alternative strategies that can be used by analysts to properly
21 account for knowledge or lack of knowledge about correlations and dependencies among the
22 variables in a risk assessment. These strategies range from functionally modeling the with
23 understanding of the mechanistic relationships between the variables, building models with
24 conditional distributions that reflect the importance dependencies, developing statistical
25 (phenomenological) models of the statistical correlation between variables, and bounding the
26 output distributions without making unjustified assumptions about the dependence between the
27 input variables.

28 **ACKNOWLEDGMENTS**

29 This work was supported in part by a [REDACTED]
30 [REDACTED] from the National Institutes of Environmental Health Sciences

1 (NIEHS), a component of the National Institutes of Health (NIH), and by contract [REDACTED] with
2 Sandia National Laboratories (SNL), under the [REDACTED]
3 [REDACTED]. Any opinions, findings, conclusions or recommendations expressed herein are solely
4 those of the authors and do not necessarily reflect the views of the NIEHS, NIH or SNL. An
5 greatly extended version of this paper can be found in Ferson et al. (2004).

6 REFERENCES

- 7 Barton W.T. (1989). Response from Palisade Corporation. *Risk Analysis* 9: 259-260.
- 8 Bedford, T.J., and R.M. Cooke. (2002). Vines —a new graphical model for dependent random
9 variables. *Annals of Statistics* 30(4): 1031-1068. Available on-line at
10 <http://ssor.twi.tudelft.nl/~risk/files/mv3.pdf>.
- 11 Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New
12 York.
- 13 Berleant, D. (1993). Automatically verified reasoning with both intervals and probability density
14 functions. *Interval Computations* 1993 (2): 48-70.
- 15 Berleant, D. and C. Goodman-Strauss. (1998). Bounding the results of arithmetic operations on
16 random variables of unknown dependency using intervals. *Reliable Computing* 4: 147-165.
- 17 Berleant, D. and J. Zhang. (2004). Using Pearson correlation to improve envelopes around the
18 distributions of functions. *Reliable Computing* 10: 139-161. Available on-line at
19 <http://class.ee.iastate.edu/berleant/home/me/cv/papers/correlationpaper.pdf>.
- 20 Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of*
21 *Mathematical Statistics* 21: 593-600.
- 22 Bratley, P., B.L. Fox and L.E. Schrage (1983). *Guide to Simulations*. Springer-Verlag, New York.
- 23 Burgman, M.A., S. Ferson and H.R. Akçakaya. (1993). *Risk Assessment in Conservation Biology*.
24 Chapman and Hall, London.
- 25 Calabrese, E.J. and C.E. Gilbert. (1993). Lack of total independence of uncertainty factors (UFs):
26 implications for the size of the total uncertainty factor. *Regulatory Toxicology and*
27 *Pharmacology* 17: 44-51.
- 28 Cario, M.C. and B. L. Nelson. (1997). Modeling and generating random vectors with arbitrary
29 marginal distributions and correlation matrix.
30 <http://www.iems.nwu.edu/%7Enelsonb/norta4.ps>
- 31 Caselton, W. F. and W. Luo (1992). Decision making with imprecise probabilities: Dempster-
32 Shafer theory and application. *Water Resources Research* 28(12): 3071–3083.

- 1 Cheng, W.-Y. (2003). Default correlations (a Microsoft Power Point presentation file).
2 <http://personal.cityu.edu.hk/~efcheng/teaching/ef5154/defaultcorrelation.ppt>.
- 3 Clemen, R. and T. Reilly (1999). Correlations and copulas for decision and risk analysis.
4 *Management Science* 45: 208-224.
- 5 Cooke, R.M. (1997). Markov and entropy properties of tree and vine-dependent variables.
6 {variant title: Markov trees and tree dependent random variables] *Proceedings of the ASA*
7 *Section on Bayesian Statistical Science*. Available on-line at
8 <http://ssor.twi.tudelft.nl/~risk/files/marktre.pdf>.
- 9 Cooke, R. (2002). *Unicorn - Uncertainty Analysis with Correlations*. Available from
10 <http://ssor.twi.tudelft.nl/~risk/software/unicorn.html>.
- 11 Cossette, H., M. Denuit and É. Marceau. (2001). Distributional bounds for functions of dependent
12 risks. Discussion paper 0128, Institut de Statistique, Université Catholique de Louvain,
13 <http://www.stat.ucl.ac.be/ISpub/dp/2001/dp0128.ps>.
- 14 Couso, I., S. Moral and P. Walley. (1999). Examples of independence for imprecise probabilities.
15 Pages 121-130 in *Proceedings of the First International Symposium on Imprecise*
16 *Probability and Their Applications*. G. de Cooman, F.F. Cozman, S. Moral and P. Walley
17 (eds.), Imprecise Probabilities Project, Universiteit Gent.
- 18 Couso, I., S. Moral and P. Walley. (2000). A survey of concepts of independence for imprecise
19 probabilities. *Risk Decision and Policy* 5: 165-181.
- 20 Cozman, F.G. and P. Walley. (2001). Graphoid properties of epistemic irrelevance and
21 independence. Pages 112-121 in *Proceedings of the Second International Symposium on*
22 *Imprecise Probability and Their Applications*. G. de Cooman, T.L. Fine and T. Seidenfeld
23 (eds.), Shaker Publishing, Maastricht.
- 24 Cullen, A.C., and H.C. Frey. (1999). *Probabilistic Techniques in Exposure Assessment: A*
25 *Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press:
26 New York.
- 27 Dall'Aglio, G., Kotz, S. and G. Salinetti (eds.) (1991). *Advances in Probability Distributions with*
28 *Given Marginals: Beyond the Copulas*, Kluwer Academic Publishers, London.
- 29 Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés: un test non
30 paramétrique d'indépendance. *Bulletin de la Classe des Sciences de l'Académie Royale de*
31 *Belgique* (ser. 5) 65: 274-292.
- 32 Decisioneering, Inc. (1996). *Crystal Ball: Forecasting and Risk Analysis for Spreadsheet Users*,
33 Aurora, Colorado.
- 34 Eves, H. (1966). *Elementary Matrix Theory*. Allyn and Bacon, Boston.

1 Ferson, S. (1994). Naive Monte Carlo methods yield dangerous underestimates of tail
2 probabilities. *Proceedings of the High Consequence Operations Safety Symposium*, Sandia
3 National Laboratories, SAND94-2364, J.A. Cooper (ed.), pp. 507-514.

4 Ferson, S. (1996a). Automated quality assurance checks on model structure in ecological risk
5 assessments. *Human and Environmental Risk Assessment* 2:558-569.

6 Ferson, S. (1996b). What Monte Carlo methods cannot do. *Human and Ecological Risk*
7 *Assessment* 2:990-1007.

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

10 Ferson, S., J. Hajagos, D. Berleant, J. Zhang, W.T. Tucker, L. Ginzburg and W. Oberkampf.
11 (2004). Dependence in Dempster-Shafer theory and probability bounds analysis. Sandia
12 National Laboratories, Technical Report SAND2002-xxxx [to appear], Albuquerque, New
13 Mexico. Available on-line at <http://www.ramas.com/d.pdf>.

14 Fetz, T. (2001). Sets of joint probability measures generated by weighted marginal focal sets.
15 Pages 171-178 in *Proceedings of the Second International Symposium on Imprecise*
16 *Probability and Their Applications*. G. de Cooman, T.L. Fine and T. Seidenfeld (eds.),
17 Shaker Publishing, Maastricht.

18 Fetz, T. and M. Oberguggenberger. (2004). Propagation of uncertainty through multivariate
19 functions in the framework of sets of probability measures. *Reliability Engineering and*
20 *System Safety* [to appear].

21 Flury, B.K. (1986). On sums of random variables and independence. *The American Statistician*
22 40: 214-215.

23 Frank, M.J., Nelsen, R.B., and Schweizer, B. (1987). Best-possible bounds for the distribution of
24 a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74, 199-211.

25 Fréchet, M. (1935). Généralisations du théorème des probabilités totales. *Fundamenta*
26 *Mathematica* 25: 379-387.

27 Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ.*
28 *Lyon, Sect. A* 9: 53-77.

29 Genest, C. and J. MacKay (1986). The joy of copulas: bivariate distributions with uniform
30 marginals. *The American Statistician* 40: 280-283.

31 Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika* 74: 549-555.

32 Goovaerts, M.J., J. Dhaene and A. De Schepper. (2000). Stochastic upper bounds for present
33 value functions. *Journal of Risk and Insurance* 67: 1-14.

- 1 Haas, C.N. (1999). On modelling correlated random variables in risk assessment. *Risk Analysis*
2 19: 1205-1214.
- 3 Hart, A., S. Ferson and J. Shaw. 2004 [tentative]. Problem formulation for probabilistic
4 ecological risk assessments. *Proceedings of the SETAC Pellston Workshop on the*
5 *Application of Uncertainty Analysis to Ecological Risks of Pesticides*. SETAC Press,
6 Pensacola, Florida.
- 7 Helton, J.C. (1993). Uncertainty and sensitivity analysis techniques for use in performance
8 assessment for radioactive waste disposal. *Reliability Engineering and System Safety* 42:
9 327-367.
- 10 Helton, J.C. and F.J. Davis. (2002). Latin hypercube sampling and the propagation of uncertainty
11 in analyses of complex systems. SAND2001-0417, Sandia National Laboratories,
12 Albuquerque, New Mexico.
- 13 Helton, J.C. and F.J. Davis. (2003). Latin hypercube sampling and the propagation of uncertainty
14 in analyses of complex systems. *Reliability Engineering and System Safety* 81: 23-69.
- 15 Henderson, S.G., B.A. Chiera and R.M. Cooke. (2000). Generating “dependent” quasi-random
16 numbers. Pages 527-536 in *Proceedings of the 2000 Winter Simulation Conference*. J. A.
17 Joines, R. R. Barton, K. Kang, and P. A. Fishwick (eds.), IEEE. Available on-line at
18 <http://ssor.twi.tudelft.nl/~risk/files/wsc2000hendersons-1i.pdf>.
- 19 Hickman, J.W., et al. (1983). PRA procedures guide: a guide to the performance of probabilistic
20 risk assessments for nuclear power plants, in two volumes. NUREG-CR-2300-V1 and -V2,
21 National Technical Information Service, Washington.
- 22 Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen*
23 *Instituts und des Instituts für Angewandte Mathematik der Universität Berlin* 5 (Heft 3): 179-
24 233 [translated in Hoeffding, W. (1940). Scale-invariant corelation theory. *Collected works*
25 *of Wassily Hoeffding*, N.I. Fisher and P.K. Sen (eds.), Springer-Verlag, New York].
- 26 Hoeffding, W. (1947). On the distribution of the rank correlation coefficient t when the variates
27 are not independent. *Biometrika* 34: 183-196.
- 28 Hutchinson, T.P. and C.D. Lai. (1990). *Continuous Bivariate Distributions, Emphasizing*
29 *Applications*. Rumsby Scientific Publishing, Adelaide 5000, Australia.
- 30 Iman, R.L. and W.J. Conover. (1980). Small sample sensitivity analysis techniques for computer
31 models, with an application to risk assessment. *Communications in Statistics A9*: 1749-1842.
- 32 Iman, R.L. and W.J. Conover. (1982). A distribution-free approach to inducing rank correlation
33 among input variables. *Communications in Statistics B11*: 311-334.

1 Iman, R.L. and J.M. Davenport (1980). Rank correlation plots for use with correlated input
2 variables in simulation studies. SAND80-1903, Sandia National Laboratories, Albuquerque.

3 Iman, R.L. and J.M. Davenport (1982a). Rank correlation plots for use with correlated input
4 variables. *Communications in Statistics (Simulation and Computation)* 11: 335-360.

5 Iman, R.L. and J.M. Davenport (1982b). An interactive algorithm to produce a positive-definite
6 correlation matrix from an approximate correlation matrix (with a program users' guide).
7 SAND81-1376, Sandia National Laboratories, Albuquerque.

8 Iman, R.L. and J.C. Helton (1985). A comparison of uncertainty and sensitivity analysis
9 techniques for computer models. NUREG/CR-3904. National Technical Information
10 Service, Springfield, Virginia.

11 Iman, R.L. and M.J. Shortencarier. (1984). *A Fortran 77 Program and User's Guide for the*
12 *Generation of Latin Hypercube and Random Samples for Use with Computer Models.*
13 Technical Report NUREG/CR-3624, SAND83-2365, Sandia National Laboratories,
14 Albuquerque, New Mexico.

15 Iman, R.L., J.C. Helton, J.E. Campbell. (1981a). An approach to sensitivity analysis of computer
16 models, Part 1. Introduction, input variable selection and preliminary variable assessment.
17 *Journal of Quality Technology* 13:174-183.

18 Iman, R.L., J.C. Helton, J.E. Campbell. (1981b). An approach to sensitivity analysis of computer
19 models, Part 2. Ranking of input variables, response surface validation, distribution effect
20 and technique synopsis. *Journal of Quality Technology* 13:232-240.

21 Insua, D.R. and F. Ruggeri (eds.) (2000). *Robust Bayesian Analysis*. Lecture Notes in Statistics,
22 Volume 152. Springer-Verlag, New York.

23 Kaas, R., J. Dhaene and M.J. Goovaerts. (2000). Upper and lower bounds for sums of random
24 variables. *Insurance: Mathematics and Economics* 27: 151-168.

25 Kowalski, C.J. (1973). Non-normal bivariate distributions with normal marginals. *The American*
26 *Statistician* 27: 103-106.

27 Kurowicka, D. and Cooke, R.M. (2002). The vine copula method for representing high
28 dimensional dependent distributions: application to continuous belief nets. Proceedings of
29 the 2002 Winter Simulation Conference E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M.
30 Charnes (eds.), IEEE. Available on-line at [http://ssor.twi.tudelft.nl/~risk/files/kurowickad-](http://ssor.twi.tudelft.nl/~risk/files/kurowickad-2i.zip)
31 [2i.zip](http://ssor.twi.tudelft.nl/~risk/files/kurowickad-2i.zip).

32 Kurowicka, D., J. Misiewicz and R.M. Cooke (2001). Elliptical copulae. Pages 209-214 in *Monte*
33 *Carlo Simulation*, Schueller and Spanos (eds), Balkema, Rotterdam. Available on-line at
34 <http://ssor.twi.tudelft.nl/~risk/files/ellcop.pdf>.

- 1 Lehmann, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* 37:
2 1137-1153.
- 3 Li, D. (2000). On default correlation: a copula function approach. *Journal of Fixed Income* 9(4):
4 43-54.
- 5 Lurie, P.M., M.S. Goldberg. (1994). A method for simulating correlated random variables from
6 partially specified distributions. Institute for Defense Analysis IDA paper P-2998,
7 Alexandria, VA.
- 8 Makarov, G.D. (1981). Estimates for the distribution function of a sum of two random variables
9 when the marginal distributions are fixed. *Theory of Probability and its Applications* 26: 803-
10 806.
- 11 Melnick, E.L. and A. Tenenbein. (1982). Misspecification of the normal distribution. *The*
12 *American Statistician* 36: 372-373.
- 13 Metzger, J. N., Fjeld, R. A., Hammonds, J. S., Hoffman, F. O. (1998). Evaluation of software for
14 propagating uncertainty through risk assessment models. *Human and Ecological Risk*
15 *Assessment* 4(2): 263-290.
- 16 Morgan, M.G. and M. Henrion (1998). Analytica: a software tool for uncertainty analysis and
17 model communication. Chapter 10 of *Uncertainty: A Guide to Dealing with Uncertainty in*
18 *Quantitative Risk and Policy Analysis*, Cambridge University Press, New York. Available
19 on-line at <http://www.lumina.com/software/ch10.9.PDF>.
- 20 Müller, A. (1997). Stop-loss order for portfolios of dependent risks. *Insurance: Mathematics and*
21 *Economics* 21: 219-223.
- 22 Nelsen, R.B. (1986). Properties of a one-parameter family of bivariate distributions with specified
23 marginals. *Communications in Statistics (Theory and Methods)* A15:3277-3285.
- 24 Nelsen, R.B. (1987). Discrete bivariate distributions with given marginals and correlation.
25 *Communications in Statistics (Simulation and Computation)* B16:199-208.
- 26 Nelsen, R.B. (1991). Copulas and association. Pages 51-75 in *Advances in Probability*
27 *Distributions with Given Marginals: Beyond the Copulas*, G. Dll'Aglio, S. Kotz and G.
28 Salinetti (eds.), Kluwer Academic Publishers, London.
- 29 Nelsen, R.B. (1995). Copulas, characterization, correlation and counterexamples. *Mathematics*
30 *Magazine* 68: 193-198.
- 31 Nelsen, R.B. (1999). *An Introduction to Copulas*. Lecture Notes in Statistics 139, Springer-
32 Verlag, New York.

- 1 Nelsen, R.B., J.J. Quesada-Molina, J.A.Rodríguez-Lallena and M. Úbeda-Flores. (2001). Bounds
2 on bivariate distribution functions with given marginals and measures of association.
3 *Communications in Statistics (Theory and Methods)* 30: 1155-1162.
- 4 Nelsen, R.B., J.J. Quesada-Molina, J.A.Rodríguez-Lallena and M. Úbeda-Flores. (2004). Best-
5 possible bounds on sets of bivariate distribution functions.*Journal of Multivariate Analyses*
6 [to appear].
- 7 Palisade Corporation. (1996). *@Risk: Advanced Risk Analysis for Spreadsheets*. Newfield, New
8 York.
- 9 Salmento, J.S., E.S. Rubin, and A.M. Finkel. (1989). A review of @Risk. *Risk Analysis* 9: 255-
10 257.
- 11 Scarsini, M. (1984). On measures of concordance. *Stochastica* 8: 201-218.
- 12 Scheuer, E.M. and D.S. Stoller. (1962). On the generation of normal random vectors.
13 *Technometrics* 4:278-281.
- 14 Schweizer, B. (1991). Thirty years of copulas.*Advances in Probability Distributions with*
15 *Given Marginals: Beyond the Copulas*, G. Dll’Agllo, S. Kotz and G. Salinetti (eds.),
16 Kluwer Academic Publishers, London.
- 17 Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.
- 18 Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de*
19 *L’Institut de Statistiques de l’Université de Paris* 8: 229-231.
- 20 Smith, A.E., P.B. Ryan and J.S. Evans. (1992). The effect of neglecting correlations when
21 propagating uncertainty and estimating the population distribution of risk. *Risk Analysis*
22 12:467-474.
- 23 Song, P. (1997). Generating dependent random numbers with given correlations and
24 margins from exponential dispersion models. *Journal of Statistical Computing and*
25 *Simulation* 56: 317-335.
- 26 Spearman, C. (1904). “General intelligence”, objectively determined and measured. *American*
27 *Journal of Psychology* 15: 201-293. See a republication at
28 <http://psychclassics.yorku.ca/Spearman/>.
- 29 Tonon, F., A. Bernadini and A. Mammino. (2000a). Reliability analysis of rock mass response
30 by means of Random Set Theory. *Reliability Engineering and System Safety* 70: 263-282.
- 31 Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London.
- 32 Wartenberg, D., S. Ferson, and F.J. Rohlf. (1987). Putting things in order: a critique of detrended
33 correspondence analysis. *The American Naturalist* 129: 434-448.

1 Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of Statistics* 4:1280-
2 1289.

3 Williamson, R.C. and T. Downs. (1990). Probabilistic arithmetic I: Numerical methods for
4 calculating convolutions and dependency bounds. *International Journal of Approximate*
5 *Reasoning* 4 89-158.

6