

Practical Probability with Spreadsheets

Chapter 4: CORRELATION OF RANDOM VARIABLES

1. Joint distributions of discrete random variables

Consider a young investor who owns some stock shares in two companies. She plans to sell these shares next year, for a down payment on a condominium. The price per share at which she will be able to sell these stocks next year is now an unknown quantity. So she begins to think about probability distributions for these unknown quantities. For notation, let \mathbf{X} denote next year's price per share of the first company that she has invested in, and let \mathbf{Y} denote next year's price per share of the second company. Suppose that she assesses the following simple discrete probability distributions for these two unknown quantities:

x	$P(\mathbf{X}=x)$	y	$P(\mathbf{Y}=y)$
70	0.2	30	0.3
75	0.5	32	0.6
80	0.3	34	0.1

Let \mathbf{W} denote the total value that she will get from selling her portfolio next year. Of course \mathbf{W} is also unknown now, because it depends on these unknown prices \mathbf{X} and \mathbf{Y} . To be specific, suppose that she owns 100 shares of the stock that will worth \mathbf{X} and she owns 200 shares of the stock that will be worth \mathbf{Y} . Then the total value of her portfolio \mathbf{W} is given by the formula

$$\mathbf{W} = 100*\mathbf{X} + 200*\mathbf{Y}.$$

With this formula and the investor's assessed probability distribution for \mathbf{X} and \mathbf{Y} , can we compute the probability distribution for \mathbf{W} ? Unfortunately, the answer is No. For example, try to compute the probability that her portfolio will be worth $\mathbf{W} = \$13000$, which (with the possible values of \mathbf{X} and \mathbf{Y} listed above) can occur only if $\mathbf{X}=70$ and $\mathbf{Y}=30$. We know that $P(\mathbf{X}=70)$ is 0.2, and $P(\mathbf{Y}=30)$ is 0.3, but what is the probability of getting $\mathbf{X}=70$ and $\mathbf{Y}=30$ at the same time? The answer to that question depends on what the investor believes about the relationship between \mathbf{X} and \mathbf{Y} . If she believes that \mathbf{X} tends to be low when \mathbf{Y} is low then the probability of getting both $\mathbf{X}=70$ and $\mathbf{Y}=30$ could be as much as 0.2. But if she believes that \mathbf{X} will be low only when \mathbf{Y} is not low, then the probability of getting both $\mathbf{X}=70$ and $\mathbf{Y}=30$ could be 0.

Mathematicians use the symbol " \cap " to denote the intersection of two events, that is, the event that

two other events both occur. Unfortunately, the symbol " \cap " cannot be used Excel labels, so we will use the ampersand symbol "&" instead of " \cap " here to denote intersections. Thus, the expression " $X=70 \& Y=30$ " is used to denote the event that X is 70 and Y is 30 (so that $W=100*70+200*30 = 13000$). With this notation, we may say that our problem above is that knowing $P(X=70)$ and $P(Y=30)$ does not tell us enough to compute $P(X=70 \& Y=30)$.

A table or function that specifies, for each pair of possible values (x,y) , the probability $P(X=x \& Y=y)$ is called the joint probability distribution of X and Y . To compute the probability distribution, we need to assess the joint probability distribution of X and Y .

When X and Y are independent random variables, then their joint probability distribution can be computed by the simple formula

$$P(X=x \& Y=y) = P(X=x) * P(Y=y).$$

So for this example, if X and Y are independent, then $P(X=70 \& Y=30)$ is $0.2*0.3 = 0.06$. With this multiplicative formula, specifying the separate distributions of independent random variables implicitly gives us their joint distribution. That is why, in Chapter 3, we did not worry about joint probability distributions of the multiple unknown quantities that arose in the "Superior Semiconductors (B)," because we were assuming that these unknown quantities were all independent of each other.

But stock prices often move up and down together, following general market trends. Thus, there is good reason to suppose that our investor might not think of X and Y as independent random variables. The discrete distributions shown above allow nine possible pairs of values for X and Y together

$$\{(70,30), (70,32), (70,34), (75,30), (75,32), (75,34), (80,30), (80,32), (80,34)\}$$

and we may ask our investor to subjectively assess a probability of each pair. She may have some difficulty assessing these nine probabilities (which must sum to 1 because exactly one pair will occur), but let us suppose that her answers are as summarized in the joint probability table shown in the range A3:D6 of Figure 1.

	A	B	C	D	E	F
1	Joint probabilities $P(X=x\&Y=y)$					
2		y=				
3	x= \	30	32	34		P(X=x)
4	70	0.1	0.1	0		0.2
5	75	0.2	0.3	0		0.5
6	80	0	0.2	0.1		0.3
7		P(Y=y)				sum
8		0.3	0.6	0.1		1
9						
10	X	Y				
11	75.5	31.6	Expected Value			
12	3.5	1.2	Standard deviation			
13						
14	FORMULAS FROM RANGE A1:F12					
15	F4.	=SUM(B4:D4)				
16	F4 copied to F4:F6.					
17	B8.	=SUM(B4:B6)				
18	B8 copied to B8:D8.					
19	F8.	=SUM(B4:D6)				
20	A11.	=SUMPRODUCT(A4:A6,F4:F6)				
21	B11.	=SUMPRODUCT(B3:D3,B8:D8)				
22	A12.	=STDEVPR(A4:A6,F4:F6)				
23	B12.	=STDEVPR(B3:D3,B8:D8)				

Figure 1. Joint probability distribution of two random variables.

The separate probability distributions for **X** and **Y** can be computed from this joint probability distributions by summing across each row (to get the probability distribution of **X**) or down each column (to get the probability distribution of **Y**). These computations are performed in cells F4:F6 and B8:D8 of Figure 1, where they yield the probability distributions for **X** and **Y** that we described above. (So this joint distribution is consistent with the earlier assessment of the separate probability distributions of **X** and **Y**.) The probability distribution of a single random variable is sometimes called its marginal probability distribution when it is computed from a joint distribution in this way, so called simply because it is often displayed next to the margin of the joint probability table, as in this spreadsheet.

In the joint probability distribution table, as shown in Figure 1, we can see more probability in the top-left and bottom-right corners, where both unknowns are greater than the

expected value or both are less than the expected value, than in the top-right and bottom-left corners, where one unknown is greater than expected and the other is less than expected. Thus, the joint probability table in Figure 1 clearly exhibits a kind of positive relationship between **X** and **Y**, as stock prices that tend to go up or down together. To provide a quantitative measure of such relationships or tendencies of random variables to co-vary, mathematicians have defined two concepts called covariance and correlation.

But before introducing these concepts, it will be helpful to switch to a different way of tabulating joint probability distributions, because the two-dimensional table shown in Figure 1 is not so convenient for computations in spreadsheets. A more convenient (if less intuitive) way of displaying a joint probability distribution is to list each possible combination of values of **X** and **Y** in a separate row, with all the joint probabilities listed down one column. Such a one-dimensional joint probability table is shown in the range A4:C13 of Figure 2. The nine possible pairs of values of **X** and **Y** are listed in rows 5 through 13, with **X** values in A5:A13 and **Y** values in B5:B13, and the probabilities for these (**X,Y**) pairs are listed in cells C5:C13. (A Simtools procedure called "Combine Rows" can be used for automatically generating all possible combinations of two lists for a range like A5:B13 here, but it is also easy to just type them in.) You should verify that the joint probabilities shown in cells C5:C13 of Figure 2 are really the same as those shown in cells B4:D6 of Figure 1.

Cells I25:J25 in Figure 2 show how to make a simulation of **X** and **Y** with this joint probability distribution. The formulas in I25 and J25 are respectively

=DISCRINV(\$I\$24,A5:A13,\$C\$5:\$C\$13)

=DISCRINV(\$I\$24,B5:B13,\$C\$5:\$C\$13)

where cell I24 contains the formula =RAND(). Because the same RAND in I24 is used in both of these formulas, the two DISCRINV functions select the same row in the C5:C13 probabilities (the lowest row above which the sum of probabilities is less than I24) and then return the values for **X** and **Y** from the A and B cells in that row. If we had used different RANDs (instead of I24) in these formulas, then cells I25 and I26 would have been independent random variables which could (for example) have taken the values 70 and 34 at the same time, even though the event **X=70&Y=34** has zero probability in the joint distribution.

	A	B	C	D	E	F	G	H	I	J	K
1							Portfolio shares				
2							of X	of Y			
3			Joint Probys				100	200			
4	x	y	P(X=x&Y=y)		ProdDevsFromMean			W = Portfolio value			
5	70	30	0.1		8.8				13000		
6	70	32	0.1		-2.2				13400		
7	70	34	0		-13.2				13800		
8	75	30	0.2		0.8				13500		
9	75	32	0.3		-0.2				13900		
10	75	34	0		-1.2				14300		
11	80	30	0		-7.2				14000		
12	80	32	0.2		1.8				14400		
13	80	34	0.1	sum	10.8				14800		
14				1							
15	X	Y		Covariance		Correlation			W	=100*X+200*Y	
16	75.5	31.6	E		2.2	0.524			13870	E	13870
17	3.5	1.2	Stdev		2.2	0.524			517.78	Stdev	517.78
18									268100	Var	268100
19	PRODS(Stdevs)			CorrelArray			PRODS(Shares)				
20	12.25	4.2		1	0.524		10000	20000			
21	4.2	1.44		0.524	1		20000	40000			
22											
23				CovarArray			Simulation model				
24				12.25	2.2				0.2167	(rand)	
25				2.2	1.44				X	Y	W
26									75	30	13500
27	FORMULAS										
28	E5. =(A5-\$A\$16)*(B5-\$B\$16)						E5 copied to E5:E13.				
29	I5. =SUMPRODUCT(\$G\$3:\$H\$3,A5:B5)						I5 copied to I5:I13.				
30	D14. =SUM(C5:C13)										
31	A16. =SUMPRODUCT(A5:A13,\$C\$5:\$C\$13)						A16 copied to B16,E16,I16.				
32	A17. =STDEVPR(A5:A13,\$C\$5:\$C\$13)						A17 copied to B17,I17.				
33	F16. =E16/(A17*B17)										
34	K16. =SUMPRODUCT(G3:H3,A16:B16)										
35	E17. =COVARPR(A5:A13,B5:B13,C5:C13)										
36	F17. =CORRELPR(A5:A13,B5:B13,C5:C13)										
37	I18. =I17^2										
38	A20:B21. {=PRODS(A17:B17)}										
39	E20,D21. =F16										
40	G20:H21. {=PRODS(G3:H3)}										
41	D24. =A20*D20						D24 copied to D24:E25.				
42	K18. =SUMPRODUCT(A20:B21,D20:E21,G20:H21)										
43	K17. =SUMPRODUCT(PRODS(A17:B17),D20:E21,PRODS(G3:H3))^0.5										
44	I24. =RAND()										
45	I26. =DISCRINV(\$I\$24,A5:A13,\$C\$5:\$C\$13)										
46	J26. =DISCRINV(\$I\$24,B5:B13,\$C\$5:\$C\$13)										
47	K26. =SUMPRODUCT(G3:H3,I26:J26)										

Figure 2. Correlation of two discrete random variables and portfolio analysis.

2. Covariance and correlation

In general, the covariance of any two random variables \mathbf{X} and \mathbf{Y} is defined to be the expected value of the product of their deviations from their respective means. That is,

$$(1) \quad \text{Covar}(\mathbf{X}, \mathbf{Y}) = E((\mathbf{X} - \mu_x) * (\mathbf{Y} - \mu_y)).$$

For our example, the mean μ_x is computed in cell A16 of Figure 2 by the formula

$$=\text{SUMPRODUCT}(A5:A13, \$C\$5:\$C\$13)$$

and the mean μ_y is computed in cell B16 by copying from cell A16. Then the deviations from the means $(\mathbf{X} - \mu_x) * (\mathbf{Y} - \mu_y)$ are computed for the nine possible pairs of (\mathbf{X}, \mathbf{Y}) values by entering the formula

$$=(A5 - \$A\$16) * (B5 - \$B\$16)$$

into cell E5 of Figure 2, and then copying E5 to E5:E13. Now to compute the expected product of deviations from the means, we multiply each possible product of deviations by its corresponding probability and sum. So the covariance can be computed by the formula

$$=\text{SUMPRODUCT}(E5:E13, \$C\$5:\$C\$13)$$

in cell E16 of Figure 2.

How should we interpret this covariance? Recall first that the product of two numbers is negative only when one is negative and one is positive; the product of two negative numbers is positive. So the product of deviations from means is positive in the cases the two random variables deviate from their respective means in the same direction (that is, when both are greater than their means or both are less than their means). The products of deviations from means is negative only when the two random variables are deviating from their respective means in opposite directions (one greater than its mean while the other is less). So the covariance is a positive number here because there is more probability in the cases where the two random variables are varying from their means in the same direction, and there is less probability in the cases where the two random variables are varying from their means in opposite directions. In this sense, the covariance is a measure of the tendency of two random variables to co-vary.

Notice also that the covariance of a random variable with itself is the random variables variance, which is the square of the standard deviation. That is,

$$\text{Covar}(\mathbf{X}, \mathbf{X}) = E((\mathbf{X} - \mu_x) * (\mathbf{X} - \mu_x)) = E((\mathbf{X} - \mu_x)^2) = \text{Var}(\mathbf{X}) = \text{Stdev}(\mathbf{X})^2.$$

(Recall that the variance of \mathbf{X} was defined as the expected value of the squared deviation of \mathbf{X} from its own expected value, and the standard deviation was defined as the square root of the variance. Here "Var" stands for "variance," but "VaR" stands for "value at risk.")

The actual numerical value of the covariance (2.2 in this example) is hard to interpret because (like the variance) it has a very strange unit of measurement. When our random variables \mathbf{X} and \mathbf{Y} are measured in dollars, their means and standard deviations are also measured in dollars. But the products of deviations involve multiplying dollar amounts by other dollar amounts, and so the covariance is measured in dollars times dollars or dollars-squared. (If we had measured the \mathbf{X} price in yen instead of dollars, keeping \mathbf{Y} in dollars, then the units of the covariance would be even stranger: dollars-times-yen!)

But when we divide the covariance of any two random variable by their standard deviations, we get a unit-free number called the correlation coefficient, which will be easier to interpret. That is, the correlation of any two random variables \mathbf{X} and \mathbf{Y} is defined by the formula

$$(2) \quad \text{Correl}(\mathbf{X}, \mathbf{Y}) = \text{Covar}(\mathbf{X}, \mathbf{Y}) / (\text{Stdev}(\mathbf{X}) * \text{Stdev}(\mathbf{Y}))$$

For our example, the correlation of \mathbf{X} and \mathbf{Y} is computed by the formula =E17/(A17*B17) in cell F17 in Figure 2.

(Simtools gives us special functions COVARPR and CORRELPR for computing the covariance and correlation coefficient directly from this kind of joint probability table. These functions are illustrated by the formulas =COVARPR(A5:A13,B5:B13,C5:C13) and =CORRELPR(A5:A13,B5:B13,C5:C13) in cell E17 and F17 of Figure 2, where they return the same values that we got in cells E16 and F16.)

To interpret the correlation of two random variables, you should know several basic facts. The correlation of any two random variables is always a number between -1 and 1. If \mathbf{X} and \mathbf{Y} are independent random variables then their correlation is 0. On the other hand, an extreme correlation of 1 or -1 occurs only when \mathbf{X} depends on \mathbf{Y} by a simple linear formula of the form $\mathbf{X} = c * \mathbf{Y} + d$ (where c and d are nonrandom numbers). Such a linear relation yields correlation 1 when the constant c is positive, so that \mathbf{X} increases linearly as \mathbf{Y} increases; but it yields correlation -1 when the constant c is negative, so that \mathbf{X} decreases linearly as \mathbf{Y} increases.

So when we find that \mathbf{X} and \mathbf{Y} have a correlation coefficient equal to 0.524 in this

example, we may infer the \mathbf{X} and \mathbf{Y} have a relationship that is somewhere about mid-way between independence (correlation = 0) and a perfect positively-sloped linear relationship (correlation = 1). A higher correlation would imply a relationship that looks more like a perfect linear relationship, while a smaller positive correlation would imply a relationship that looks more like independence.

3. Linear functions of several random variables

We introduced this chapter with an investor who owns 100 shares of the stock that will have price \mathbf{X} per share, and 200 shares of the stock that will have price \mathbf{Y} per share. So the total value of her portfolio \mathbf{W} will depend on \mathbf{X} and \mathbf{Y} by the linear formula $\mathbf{W} = 100*\mathbf{X} + 200*\mathbf{Y}$. In Figure 2, the numbers of shares that the investor owns in these two stocks have been entered into cells G3 and H3, and the possible portfolio values corresponding to each possible (\mathbf{X}, \mathbf{Y}) pair are shown in column I. To compute the value of \mathbf{W} that corresponds to the (\mathbf{X}, \mathbf{Y}) values in cells A5:B5, cell I5 contains the formula

$$=\text{SUMPRODUCT}(\$G\$3:\$H\$3,A5:B5),$$

and this formula from cell I5 has been copied to the range I5:I13. Then $E(\mathbf{W})$ and $\text{Stdev}(\mathbf{W})$ can be computed by the formulas

$$=\text{SUMPRODUCT}(I5:I13,\$C5:C13)$$

$$=\text{STDEVPR}(I5:I13,\$C5:C13)$$

in cells I16 and I17 respectively. The variance $\text{Var}(\mathbf{W})$, as the square of the standard deviation, is computed in cell I18 of Figure 2 by the formula

$$=I17^2$$

But the expected value and variance of \mathbf{W} can also be calculated another way, by using some general formulas about the means and variances of linear functions of random variables. These formulas are complicated, but I can state them for you concisely in the next two paragraphs, and then we begin to explore their significance in statistics and finance.

Given any number n , suppose that $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are a collection of n random variables. Consider any random variable \mathbf{A} that is defined by the linear functions:

$$\mathbf{A} = \alpha_0 + \sum_{i=1}^n \alpha_i * \mathbf{X}_i$$

where α_i are nonrandom numbers for each i in $\{0,1,\dots,n\}$. Then

$$(3) \quad E(\mathbf{A}) = \alpha_0 + \sum_{i=1}^n \alpha_i * E(\mathbf{X}_i),$$

$$(4) \quad \text{Var}(\mathbf{A}) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i * \alpha_j * \text{Covar}(\mathbf{X}_i, \mathbf{X}_j)$$

To express equation (4) in terms of standard deviations and correlations, recall that $\text{Covar}(\mathbf{X}_i, \mathbf{X}_j)$ equals $\text{Correl}(\mathbf{X}_i, \mathbf{X}_j) * \text{Stdev}(\mathbf{X}_i) * \text{Stdev}(\mathbf{X}_j)$, and $\text{Stdev}(\mathbf{A})$ equals $\text{Var}(\mathbf{A})^{0.5}$. So we get

$$(5) \quad \text{Stdev}(\mathbf{A}) = \left(\sum_{i=1}^n \sum_{j=1}^n \alpha_i * \alpha_j * \text{Correl}(\mathbf{X}_i, \mathbf{X}_j) * \text{Stdev}(\mathbf{X}_i) * \text{Stdev}(\mathbf{X}_j) \right)^{0.5}$$

Now suppose that we have a second random variable that is also defined by another linear function of these random variables $\mathbf{X}_1, \dots, \mathbf{X}_n$, say

$$\mathbf{B} = \beta_0 + \sum_{i=1}^n \beta_i * \mathbf{X}_i$$

where β_i are nonrandom numbers for each i in $\{0,1,\dots,M\}$. Then the covariance of \mathbf{A} and \mathbf{B} is

$$(6) \quad \text{Covar}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i * \beta_j * \text{Covar}(\mathbf{X}_i, \mathbf{X}_j).$$

which is clearly a generalization of equation (4).

Consider the special case where the random variables $\mathbf{X}_1, \dots, \mathbf{X}_M$ are independent of each other. In this independent case, $\text{Covar}(\mathbf{X}_i, \mathbf{X}_j)$ and $\text{Correl}(\mathbf{X}_i, \mathbf{X}_j)$ are 0 whenever $i \neq j$. So equations (4) through (6) in the independent case become

$$(4i) \quad \text{Var}(\mathbf{A}) = (\alpha_1^2) * \text{Var}(\mathbf{X}_1) + \dots + (\alpha_n^2) * \text{Var}(\mathbf{X}_n)$$

$$(5i) \quad \text{Stdev}(\mathbf{A}) = ((\alpha_1 * \text{Stdev}(\mathbf{X}_1))^2 + \dots + (\alpha_n * \text{Stdev}(\mathbf{X}_n))^2)^{0.5}$$

$$(6i) \quad \text{Covar}(\mathbf{A}, \mathbf{B}) = (\alpha_1 * \beta_1) * \text{Var}(\mathbf{X}_1) + \dots + (\alpha_n * \beta_n) * \text{Var}(\mathbf{X}_n).$$

These are complicated formulas, and their importance will take much time for us to explore. For finance, formulas (3) and (5) tell us that, when we know the means, standard deviations, and correlations of a collection of asset prices, then we can compute the mean and standard deviation of any linear portfolio that is composed of these assets. To see how these computations can be done in spreadsheets, let us return to our example in Figure 2.

In this example, there are just two asset prices. So $n=2$ here, and \mathbf{X} and \mathbf{Y} here are taking

the place of \mathbf{X}_1 and \mathbf{X}_2 in the general formulas. The portfolio value $\mathbf{W} = 100*\mathbf{X}+200*\mathbf{Y}$ is taking the place of \mathbf{A} in the general formulas, and so our share-coefficients here are $\alpha_1 = 100$ and $\alpha_2 = 200$, and $\alpha_0 = 0$.

Formula (3) tells us that the expected value of \mathbf{W} can be computed by the simple linear formula $E(\mathbf{W}) = 100*E(\mathbf{X})+200*E(\mathbf{Y})$. With the share-coefficients 100 and 200 listed in G3:H3 of Figure 2, and with the expected values $E(\mathbf{X})$ and $E(\mathbf{Y})$ listed in A16:B16, this linear formula for $E(\mathbf{W})$ can be computed in cell K16 by the formula

$$=SUMPRODUCT(G3:H3,A16:B16)$$

To check, notice cell K16 and I16 are equal in Figure 2.

Using formula (5) to compute $Stdev(\mathbf{W})$ is much more complicated, because formula (5) refers to correlations $Correl(\mathbf{X}_i,\mathbf{X}_j)$ for all pairs of assets \mathbf{X}_i and \mathbf{X}_j that we are using in our portfolio. When there are n assets, we have n choices for \mathbf{X}_i and n choices for \mathbf{X}_j , and so we have a matrix of $n*n$ correlations appearing in this formula. To do computations in a spreadsheet, we will regularly list all these correlations together in a correlation array.

The correlation array for our example in Figure 2 is shown in the range D20:E21. With \mathbf{X} as our first asset price and \mathbf{Y} as our second asset price, the correlation array is

$$\begin{array}{ll} Correl(\mathbf{X},\mathbf{X}) = 1 & Correl(\mathbf{X},\mathbf{Y}) = 0.524 \\ Correl(\mathbf{Y},\mathbf{X}) = 0.524 & Correl(\mathbf{Y},\mathbf{Y}) = 1 \end{array}$$

In any correlation array, the elements on the diagonal from top-left to bottom-right are always ones, because the correlation of any random variable with itself is 1. Below this diagonal and to its right, the elements of the correlation array are always symmetric, because $Correl(\mathbf{X},\mathbf{Y})$ equals $Correl(\mathbf{Y},\mathbf{X})$.

Inside the summation in formula (5), we are told to multiply each number in the correlation array by a corresponding product of share-coefficients $\alpha_i*\alpha_j$ and by a corresponding product of standard deviations $Stdev(\mathbf{X}_i)*Stdev(\mathbf{X}_j)$. The four products of the share coefficients $\alpha_1=100$ and $\alpha_2=200$ are

$$\begin{array}{ll} 100*100 = 10,000 & 100*200 = 20,000 \\ 200*100 = 20,000 & 200*200 = 40,000 \end{array}$$

as shown in cells G20:H21 in Figure 2. The four products of the standard deviations $Stdev(\mathbf{X}) = 3.5$ and $Stdev(\mathbf{Y}) = 1.2$ are

$3.5*3.5 = 12.25$	$3.5*1.2 = 4.2$
$1.2*3.5 = 4.2$	$1.2*1.2 = 1.44$

as shown in cells A20:B21 of Figure 2.

It would not be difficult to compute all these products by entering multiplication formulas into each cell, but Simtools gives us a function called PRODS that makes it easier. But before I can introduce PRODS, I have to tell you first about array formulas in Excel.

An array formula in Excel is a formula that is shared by a whole range of cells at once. To enter an array formula, first select the desired range (say, A20:B21), then type the desired formula (say, =PRODS(A17:B17)), and finish with the special keystroke combination [Ctrl]-[Shift]-[Enter]. (That is, while holding down the [Ctrl] and [Shift] keys, press [Enter].) Now if you look at the formula for any cell in the range that you selected, you should see the array formula that you entered, within a pair of special braces { and }, which were added by Excel to indicate that this is an array formula. (Typing the braces yourself will not work; you must use [Ctrl]-[Shift]-[Enter]. From any selected cell that has an array formula, you can identify the whole range of cells that share the formula by the special keystroke [Ctrl]-/.)

If "values" denotes a range containing numbers in a column or row of a spreadsheet, then PRODS(values) returns a square array containing the results of multiplying each pair of numbers in the values range. The i'th row and j'th column of the array returned by PRODS will be the result of multiplying the i'th entry in the values range times the j'th entry in the values range. So in Figure 2, the array formula {=PRODS(A17:B17)} has been entered into the range A20:B21, to return the array of products of the standard deviations from A17 and B17. Similarly, the array formula {=PRODS(G3:H3)} has been entered into the range G20:H21, to return the array of products of the shares from G3 and H3.

The variance of the linear portfolio **W** can now be computed in cell K18 of Figure 2 by the formula

$$=SUMPRODUCT(A20:B21,D20:E21,G20:H21)$$

(Notice that Excel's SUMPRODUCT function can be applied to any number of ranges.) Here SUMPRODUCT multiplies each product of standard deviations in A20:B21 by the corresponding correlation in D20:E21 and by the corresponding product of share-coefficients in G20:H21, and

the results of all these various multiplications are summed, to return the same variance 268100 (\$\$) that we found before in cell I18.

The standard deviation can now be computed as the square root (that is, the power $^0.5$) of the variance. So the formula for Stdev(**W**) in cell K17 of Figure 2 is

$$=SUMPRODUCT(PRODS(A17:B17),D20:E21,PRODS(G3:H3))^0.5$$

Notice here that the PRODS function can be used inside a regular formula as well as in an array formula. That is, the 2-by-2 array returned by PRODS(A17:B17) can be interpreted directly in the SUMPRODUCT function, and it does not actually need to be entered into any other 2-by-2 range in the spreadsheet. The array formulas in A20:B21 and G20:H21 were introduced above only for pedagogical purposes.

As another important application of the linear formulas that have been introduced in this section, consider the case where **A** is the average of n independent random variables $\mathbf{X}_1, \dots, \mathbf{X}_n$ which have been drawn from the same probability distribution and which all have mean μ and standard deviation σ (variance σ^2). That is,

$$\mathbf{A} = (\mathbf{X}_1 + \dots + \mathbf{X}_n)/n = (1/n)*\mathbf{X}_1 + \dots + (1/n)*\mathbf{X}_n.$$

With this linear function, equation (3) and (4i) give us:

$$E(\mathbf{A}) = (1/n)*E(\mathbf{X}_1) + \dots + (1/n)*E(\mathbf{X}_n) = n*(1/n)*\mu = \mu$$

$$\begin{aligned} \text{Var}(\mathbf{A}) &= ((1/n)^2)*\text{Var}(\mathbf{X}_1) + \dots + ((1/n)^2)*\text{Var}(\mathbf{X}_n) = n*((1/n)^2)*(\sigma^2) \\ &= (\sigma^2)/n \end{aligned}$$

and so

$$\text{Stdev}(\mathbf{A}) = \sigma/(n^0.5).$$

Thus, we have derived the formulas for the expected value and standard deviation of a sample average that we used in Chapter 2 for computing 95% confidence intervals.

4. Multivariate Normal random variables

The Multivariate-Normal distributions are a family of joint probability distributions for collections of two or more random variables. To specify a Multivariate-Normal distribution some collection of random variables $\mathbf{X}_1, \dots, \mathbf{X}_n$, we must specify the mean $E(\mathbf{X}_i)$ and standard deviation $\text{Stdev}(\mathbf{X}_i)$ for each of these random variables, and the correlation $\text{Correl}(\mathbf{X}_i, \mathbf{X}_j)$ for each pair of

these random variables.

Suppose that $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ are Multivariate-Normal random variables; that is, their joint probability distribution is in this Multivariate-Normal family. Then each of these random variables \mathbf{X}_i is (by itself) a Normal random variable. But more importantly, any linear function of these random variables will also be a Normal random variable. That is, given any nonrandom numbers $\alpha_0, \alpha_1, \dots, \alpha_n$, the linear function

$$\mathbf{A} = \alpha_0 + \alpha_1 * \mathbf{X}_1 + \dots + \alpha_n * \mathbf{X}_n$$

will also have a Normal probability distribution.

This fact, that every linear function of Multivariate-Normals is also Normal, makes the Multivariate-Normal joint distribution very convenient for financial analysis. We have seen how the mean and standard deviation of a portfolio's value can be computed from the means, standard deviations, and correlations of the various asset prices (using formulas (3) and (5)). So if we can assume that these asset prices have a Multivariate-Normal joint distribution, then our portfolio's value will have the Normal distribution with this mean and standard deviation, and so we can easily compute its probability of being greater or less than any given number.

Any collection of independent Normal random variables is itself a collection of Multivariate-Normal random variables (with correlations that are zero for all distinct pairs). So any linear function of independent Normal random variables is also Normal. When we compute several different linear functions of the same independent Normal random variables, the results of these linear functions are Multivariate-Normal random variables that have means, standard deviations, and correlations that can be computed using formulas (3), (5i), and (6i) above.

	A	B	C	D	E	F
1	MAKING MULTIVARIATE-NORMAL RANDOM VARIABLES					
2			Constants			
3			10	20	30	
4	Independent (0,1)-Normals		Linear coefficients			
5	-0.566440121		5	7	0	
6	0.175441528		3	4	-4	
7	0.212257873		6	0	9	
8	0.201605417		-2	5	2	
9	Multivariate-Normal random variables					
10			X	Y	Z	
11			8.56446	17.74471	31.61177	
12						
13	Expected values		10	20	30	
14	Standard deviations		8.602325	9.486833	10.04988	
15						
16	Correlation array		X	Y	Z	
17		X	1	0.453382	0.439549	
18		Y	0.453382	1	-0.06293	
19		Z	0.439549	-0.06293	1	
20						
21	Any linear function of Multivariate-Normals is also Normal.					
22		Constant	1000			
23		Coefficients	100	300	200	
24	W = LinearFunction(X,Y,Z)		13502.21			
25		E(W)	14000			
26		Stdev(W)	3987.48			
27						
28	Making similar Multivariate-Normals with CORAND:					
29		CORANDs	0.296995	0.486688	0.878929	
30	Multivariate-Normal random variables					
31			X	Y	Z	
32			5.414411	19.68339	41.75485	
33	FORMULAS					
34	A5. =NORMINV(RAND(),0,1)					
35	A5 copied to A5:A8					
36	C11. =C3+SUMPRODUCT(C5:C8,\$A\$5:\$A\$8)					
37	C13. =C3					
38	C14. =SUMPRODUCT(C5:C8,C5:C8)^0.5					
39	C11:C14 copied to C11:E11					
40	D17. =SUMPRODUCT(C5:C8,D5:D8)/(C14*D14)				C18. =D17	
41	E17. =SUMPRODUCT(C5:C8,E5:E8)/(C14*E14)				C19. =E17	
42	E18. =SUMPRODUCT(D5:D8,E5:E8)/(D14*E14)				D19. =E18	
43	C24. =C22+SUMPRODUCT(C23:E23,C11:E11)					
44	C25. =C22+SUMPRODUCT(C23:E23,C13:E13)					
45	C26. =SUMPRODUCT(PRODS(C23:E23),PRODS(C14:E14),C17:E19)^0.5					
46	C29:E29. {=CORAND(C17:E19)}					
47	C32. =NORMINV(C29,C13,C14)					
48	C32 copied to C32:E32					

Figure 3. Making Multivariate-Normal random variables.

This method of creating Multivariate-Normal random variables is illustrated in Figure 3. The formula `=NORMINV(RAND(),0,1)` copied into cells A5:A8 first gives us four independent Normal random variables, each with mean 0 and standard deviation 1. Entering the formula `=C3+SUMPRODUCT(C5:C8,A5:A8)` into cell C11, and copying C11 to C11:E11, we get a collection of three nonindependent Multivariate-Normal random variables in cells C11:E11, each generated by a different linear function of the independent Normals in A5:A8.

The expected values for the three Multivariate-Normals in cells C11:E11 are just the constant terms from C3:E3, because the underlying independent Normals in A5:A8 all have expected value 0. The standard deviations for the three Multivariate-Normals in C11:E11 are computed in cells C14:E14 using formula (5i) (simplified by the fact that each of the independent Normals in A5:A8 has standard deviation 1). The correlations among the three Multivariate-Normals in C11:E11 are computed in cells D17, E17, and E18, using formulas (2) and (6i). Notice that the random variables in cells C11 and D11 have a positive correlation (0.453) mainly because C11 and D11 both depend with positive coefficients on the independent Normals in cells A5 and A6, so that any change in A5 or A6 would push C11 and D11 in the same direction. But the coefficients in cells C8 and D8 have different signs, and so any change in cell A8 would push C11 and D11 in opposite directions, which decreases the correlation of C11 and D11. The correlation between cells D11 and E11 is negative (-0.0629) mainly because the random cell A6 has opposite effects on D11 and E11, due to the coefficients 4 and -4 in cells D6 and E6.

Let me now take up a more important question. Given any possible means, standard deviations, and correlations, how can we create Multivariate-Normal random variables that fit these parameters? For example, suppose you were asked to construct three random variables that have a Multivariate-Normals joint distribution, with the means shown in C13:E13 of Figure 3, the standard deviations shown in C14:E14, and the correlations shown in C17:E19. Of course we have seen such Multivariate-Normals constructed in C11:E11, using some rather arbitrary coefficients that appear in C5:E8, but it is difficult to see how we might have such coefficients if we only knew the correlations that we wanted to implement! To avoid this difficult problem, we can use a Simtools function called CORAND, that has been designed to make it easy to simulate Multivariate-Normals with any possible correlation array.

Consider first how we would generate three Normal random variables that have the means shown in C13:E13 and the standard deviations shown in C14:E14 (ignoring for now the correlations in C17:E19). We could simply enter the formula =NORMINV(RAND(),C13,C14) into (say) cell C32, and then copy cell C32 to C32:E32. The result would be three Normal random variables that have the means and standard deviations that we wanted. But they would be independent random variables, with correlations equal to zero, because they are being driven by three different RANDs which are always independent in Excel. So we need to replace these three RANDs by random values which separately have the same Uniform distribution as a RAND but which are not independent. CORAND is designed to return such correlated random values to a range of cells in a row, when it is entered as an array formula.

The required parameter for the CORAND function is a correlation array. If the correlation array is a square with n rows and n columns, then CORAND generates a row of n random values. Each of these CORAND values is drawn from a Uniform probability distribution on the interval from 0 to 1, just like the value of a RAND function. But the values in a CORAND array are correlated precisely so that, when these CORAND values are used as the first random parameter of NORMINV functions, the resulting NORMINV values will be Multivariate-Normal random variables that have correlations as specified in the given correlation array.

The use of CORAND is illustrated in rows 29 and 32 of Figure 3. The array formula

$$\{=\text{CORAND}(\text{C17:E19})\}$$

has been entered (with the special [Ctrl]-[Shift]-[Enter] keystrokes) into the range C29:E29.

Then the formula

$$=\text{NORMINV}(\text{C29},\text{C13},\text{C14})$$

has been entered into cell C32, and C32 has been copied to C32:E32. The result is that the three cells C32, D32, and E32 contain Multivariate-Normal random variables. The means of these three Multivariate-Normal random variables are respectively C13, D13, and E13, as specified by the second parameter of their NORMINV functions; and their standard deviations are respectively C14, D14, and E14, as specified by the third parameter of their NORMINV functions. But their correlations are as specified in C17:E19, because that is the correlation array of the CORANDs that these NORMINVs use for their first inputs. To be specific, the correlation of C32 and D32

(the first and second of these Multivariate-Normals) is the value 0.453 from cell D17 (in the first row and second column of the correlation array). The correlation of C32 and E32 (the first and third of these Multivariate-Normals) is the value 0.439 from cell E17 (in the first row and third column of the correlation array). And the correlation of D32 and E32 (the second and third of these Multivariate-Normals) is the value -0.0629 from cell E18 (in the second row and third column of the correlation array).

So cells C32:E32 in Figure 3 contain random variables that have the a Multivariate-Normal joint probability distribution that has the same means, standard deviations, and correlations as the random variables above in cells C11:E11. Thus, simulation data from either range would be statistically indistinguishable from the other range, because they have exactly the same joint probability distribution.

Any possible correlation array can be implemented in this way, using CORAND. But you should know that some square arrays may not be "possible" as correlation arrays. We have already seen that a correlation array must have all 1s on the diagonal from top-left to bottom right, and there must be a symmetry between the values to the right of this diagonal and below this diagonal. But there can also be other more subtle restrictions. For example, if you tried to change the spreadsheet shown in Figure 3 by entering the value -0.9 into cell E18, then the `{=CORAND(C17:E19)}` array formula would return an error message. The problem is that you would be asking CORAND to make three random variables in which the first has a substantial positive correlation with both the second and the third, while the second and third have very strong negative correlation. This is impossible, in effect, because the first random variable cannot be kept close enough to both of the others when they are being kept far away from each other. So there can be subtle mathematical constraints on the correlations of three or more random variables. But fortunately, these feasibility constraints will always be satisfied by the correlation arrays that we estimate from statistical data.

5. Estimating correlations from data

Excel provides a statistical function CORREL for estimating the correlations among random variables that are drawn from some joint distribution which we do not know but from

which we have observed repeated independent draws. To illustrate the use of this statistical function, Figure 4 shows price data from the decade of the 1980s for six large stock funds that are tied to major market indexes. (Funds 1, 2, 3, and 4 represent broad portfolios of stocks in the New York Stock Exchange. Fund 1 contains stocks of companies that are broadly characterized as industrial, Fund 2 contains stocks of transportation companies, Fund 3 contains stocks of utility companies, and Fund 4 contains stocks of financial companies. Funds 5 and 6 are well-known selected portfolios of leading industrial companies.) The end-of-year average price-per-share data for these stock funds is shown in the range B3:G13 of Figure 4. Then the annual growth ratio for the values of these funds is then computed in range B17:G26, by entering the formula $=B4/B3$ in cell B17 and then copying cell B17 to B17:G26. So the 1980 Fund 1 growth ratio 1.22 in cell B17, for example, tells us that each dollar invested in Fund 1 just before 1980 began would have been worth \$1.22 at the end of 1980.

Our goal is to use this growth-ratio data from the 1980s to build simple forecasting models that could help to make predictions about the possible returns to various investment portfolios in the future years of the 1990s. The basic assumption of our analysis will be that the annual growth ratios of these funds are jointly drawn each year out of some joint probability distribution, which we can try to estimate from the past data. Then we can use this probability distribution to make a simulation model of future returns.

In the previous chapter, we said that the Lognormal distributions are often good fits for the annual growth ratios of financial assets. But in this chapter, we have also discussed how the Multivariate-Normal assumption would be more convenient for making predictions about the growths of different portfolios of investments in these six funds. Fortunately, our statistically estimated mean of each of these growth ratios will all be more than five times the standard deviation (see rows 14 and 15 in Figure 7 below), which makes the Lognormal and Normal distributions almost the same. So let us assume here the annual growth ratios for these six funds are jointly drawn each year out of some Multivariate-Normal probability distribution, with each year's growth ratios being another independent draw out of this joint distribution. Our statistical problem is to estimate the unknown parameters of this Multivariate-Normal distribution: the means and standard deviations for each fund and the correlations for each pair of funds.

	A	B	C	D	E	F	G	
1	End-of-year price data for six stock funds							
2		Fund 1	Fund 2	Fund 3	Fund 4	Fund 5	Fund 6	
3	1979	64.76	47.34	38.20	61.42	84.44	103.01	
4	1980	78.70	60.61	37.35	64.25	89.14	118.78	
5	1981	85.44	72.61	38.91	73.52	93.29	128.05	
6	1982	78.18	60.41	39.75	71.99	88.44	119.71	
7	1983	107.45	89.36	47.00	95.34	119.03	160.41	
8	1984	108.01	85.63	46.44	89.28	117.85	160.46	
9	1985	123.79	104.11	56.75	114.21	132.82	186.84	
10	1986	155.85	119.87	71.36	147.20	179.28	236.34	
11	1987	195.31	140.39	74.30	146.48	227.60	286.83	
12	1988	180.95	134.12	71.77	127.26	206.08	265.79	
13	1989	216.23	175.28	87.43	151.88	250.89	322.84	
14								
15	Annual growth ratios							
16		Fund 1	Fund 2	Fund 3	Fund 4	Fund 5	Fund 6	
17	1980	1.22	1.28	0.98	1.05	1.06	1.15	
18	1981	1.09	1.20	1.04	1.14	1.05	1.08	
19	1982	0.92	0.83	1.02	0.98	0.95	0.93	
20	1983	1.37	1.48	1.18	1.32	1.35	1.34	
21	1984	1.01	0.96	0.99	0.94	0.99	1.00	
22	1985	1.15	1.22	1.22	1.28	1.13	1.16	
23	1986	1.26	1.15	1.26	1.29	1.35	1.26	
24	1987	1.25	1.17	1.04	1.00	1.27	1.21	
25	1988	0.93	0.96	0.97	0.87	0.91	0.93	
26	1989	1.19	1.31	1.22	1.19	1.22	1.21	
27								
28	FORMULAS							
29	B17. =B4/B3							
30	B17 copied to B17:G26.							

Figure 4. Price data and annual growth ratios from the 1980s for six mutual funds.

We have already seen (in Chapter 2) how the unknown means and standard deviations for each random variable can be estimated from our statistical data by the sample average and the sample standard deviation, which can be computed in Excel by the functions AVERAGE and STDEV. In Figure 5, for example, the annual growth-ratio data for Fund 1 is exhibited in the range B2:B11. So cell B13 estimates the mean annual growth ratio for Fund 1 by the formula =AVERAGE(B2:B11), and cell B14 estimates the standard deviation of Fund 1's annual growth ratio by the formula =STDEV(B2:B11). Cells B18:B30 remind us of how the STDEV function computes this estimated standard deviation: by computing the squared deviation of each data

point from the sample mean, then computing an adjusted average of these squared deviations ("adjusted" in that we divide by $n-1$ instead of n), and finally returning the square root of this adjusted-average of the squared deviations.

	A	B	C	D	E	F	G	H
1		Fund 1	Fund 2		Products of deviations from means			
2	1980	1.22	1.28		0.0098			
3	1981	1.09	1.20		-0.0022			
4	1982	0.92	0.83		0.0718			
5	1983	1.37	1.48		0.0768			
6	1984	1.01	0.96		0.0260			
7	1985	1.15	1.22		0.0005			
8	1986	1.26	1.15		-0.0004			
9	1987	1.25	1.17		0.0019			
10	1988	0.93	0.96		0.0421			
11	1989	1.19	1.31		0.0087			
12								
13	Means	1.1375	1.1548		0.0261	Sum/(10-1) [Covariance]		
14	StDevns	0.1520	0.1922					
15	Correl	0.8937			0.8937	Correlation		
16								
17	Squared deviations from means				FORMULAS FROM RANGE A1:E31			
18		0.0060	0.0157		B13. =AVERAGE(B2:B11)			
19		0.0027	0.0019		C13. =AVERAGE(C2:C11)			
20		0.0495	0.1042		B14. =STDEV(B2:B11)			
21		0.0561	0.1052		C14. =STDEV(C2:C11)			
22		0.0175	0.0386		B15. =CORREL(B2:B11,C2:C11)			
23		0.0001	0.0037		B18. =(B2-B\$13)^2			
24		0.0148	0.0000		B18 copied to B18:C27			
25		0.0134	0.0003		B29. =SUM(B18:B27)/(10-1)			
26		0.0445	0.0398		C29. =SUM(C18:C27)/(10-1)			
27		0.0033	0.0231		B30. =B29^0.5	C30. =C29^0.5		
28					E2. =(B2-\$B\$13)*(C2-\$C\$13)			
29	Sum/(10-1)	0.0231	0.0370		E2 copied to E2:E11			
30	SquareRoot	0.1520	0.1922		E13. =SUM(E2:E11)/(10-1)			
31					E15. =E13/(B14*C14)			
32					B34:C34. {=CORAND(B15)}			
33	Simulation Model:				B36. =NORMINV(B34,B13,B14)			
34	CORANDs	0.4125	0.6471		C36. =NORMINV(C34,C13,C14)			
35	Multivariate Normals with same means, stdevs, and correl as above.							
36		1.10	1.23					

Figure 5. Computing the correlation of Funds 1 and 2.

The annual growth-ratio data for Funds 2 is exhibited in the range C2:C11 of Figure 5. So copying cells B13 and B14 to C13 and C14 similarly yields our statistical estimates of the mean

and standard deviation of the annual growth ratio for Fund 2.

Now to estimate the correlation between the annual growth ratios of Funds 1 and 2, the formula

$$=CORREL(B2:B11,C2:C11)$$

has been entered into cell B15 in Figure 5. The E column in Figure 5 shows in more detail how Excel's CORREL function computes this estimated correlation. First, the product of the two random variables' deviations from their respective means is computed, which is done in the spreadsheet of Figure 5 by entering the formula

$$=(B2-\$B\$13)*(C2-\$C\$13)$$

into cell E2 and then copying cell E2 to E2:E11. Next, the adjusted average (dividing by $n-1$ instead of n , where n is the number of observations that we have of each random variable) of these products of deviations is computed, as shown in cell E13 of Figure 5 by the formula

$$=SUM(E2:E11)/(10-1)$$

This adjusted-average of the observed products of deviations is a good estimate of the covariance of these two random variables (which has been defined as the expected product of their deviations from their respective means). But the correlation is the covariance divided by both of the standard deviations. So we can compute the estimated correlation from the estimated covariance in cell E15 by the formula

$$=E13/(B13*C13)$$

Notice that this formula in cell E15 returns the same value as the CORREL function in cell B15 of Figure 5.

By the law of large numbers, if we get a very large sample of paired random variables (like the growth ratios of Funds 1 and 2 here) that are drawn together from some fixed joint distribution, with each pair drawn independently of all other pairs, then the sample correlation computed by CORREL is very likely to be very close to the true correlation of the joint probability distribution from which we have sampled.

Now with CORAND and NORMINV we can easily construct Multivariate-Normal random variables that simulate the annual future growth ratios of these two funds, with the means, standard deviations, and correlation that we have estimated from our statistical data. First,

appropriately correlated RANDom values are returned in cells B34:C34 by the array formula

`{=CORAND(B15)}`

Notice that, when we want to make two correlated random variables, we only need to list their correlation coefficient (a single number) as the input parameter of CORAND. In this array formula, CORAND interprets the single number (0.8937) that it finds in cell B15 exactly as it would interpret a two-by-two correlation array with 1s in the top-left and bottom right and with this number appearing as the cross-correlation in the other two corners. Then the formulas

`=NORMINV(B34,B13,B14)` and `=NORMINV(C34,C13,C14)`

are entered into cells B36 and C36 respectively in Figure 5. The result is that cells B36:C36 contain Multivariate-Normal random variables with the means computed in B13 and C13, the standard deviations computed in B14 and C14, and with the correlation computed in B15.

Figure 6 compares the results of 200 simulations of these Multivariate-Normal random variables (on the right) with the 10 historical data points (on the left). Comparing these two scatter plots, you can see the value of having a simulation model. The model enables us to generate large simulated data sets that exhibit the same basic patterns and relationships as our real data, and thus enables us to see these patterns and relationships more clearly.

The simulation model can also give us more realistic answers to detailed questions than we might otherwise pull out of our data. Let us ask, for example, what is the probability of the event that in the coming year Fund 1's growth ratio will be less than 1 while Fund 2's growth ratio will be greater than 1. This event never happened in our 10 years of observation, but such a failure to observe the event in ten years does not prove that this event cannot happen next year. Indeed, the event "almost" occurred in 1988 for example. In the simulation data shown in Figure 6, this event occurred 7 out of 200 times. So the simulation data suggests $7/200 = 0.035$ as an estimated probability of this event, which seems more reasonable than 0.

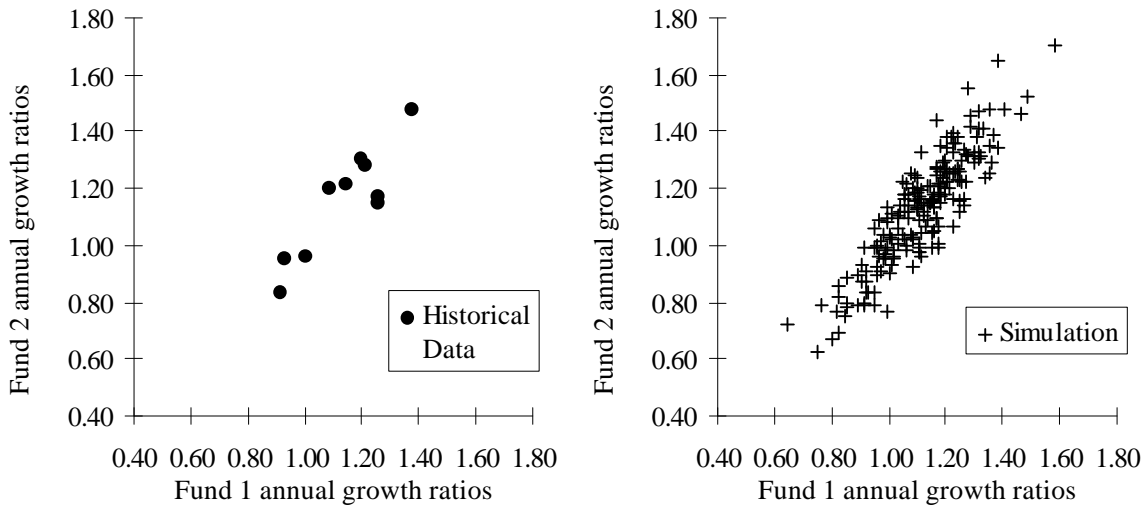


Figure 6. Extending historical data by simulation.

Figure 7 takes up the problem of estimating a Multivariate-Normal distribution for the annual growth ratios of all six funds. The observed growth ratios from Figure 4 are repeated in cells B3:G12 in Figure 7, with the ten observed growth ratios for Fund 1 in B3:B12, the ten observed growth ratios for Fund 2 in C3:C12, and so on. The expected growth ratio for each fund is estimated in row 14 by the AVERAGE of its growth data. The standard deviation of fund's growth ratio is estimated in row 5 by the sample standard deviation (STDEV) of its growth data. But with six funds, we have 15 different correlations to compute between every pair of distinct funds, to fill a six-by-six correlation array (which will be symmetric across a diagonal of 1s). Each of these pairwise correlations can be computed using Excel's CORREL, but typing in 15 different CORREL formulas can become tedious. So Simtools provides a function called MCOLL which allows us to compute all these correlations with one array formula. (The "M" in this function name stands for "matrix," which is another mathematical term for an array.)

MCORRELS returns the correlation coefficients among the columns of the data range which is specified as its input parameter. If the datarange has n columns, then MCOLL(datarange) should be entered as an array formula into an n-by-n range. The value returned by MCOLL to the i'th row and j'th column its output range is the sample correlation

that would be computed by CORREL between the i'th and j'th columns of the data range. In Figure 7, the data range is B3:G12, with one data column for each of the 6 funds. So the array formula

$$\{=MCORRELS(B3:G12)\}$$

is entered into the 6-by-6 range B19:G24 in Figure 7. (To enter this formula, I first selected the range B19:G24, then I typed the formula `=MCORRELS(B19:G24)` and then I held down the [Ctrl] and [Shift] keys while I pushed the [Enter] key.) So in the i'th row and j'th column of this range B19:G24, we find the correlation coefficient between Fund i and Fund j. For example, in Figure 7 the cell C19 is in the first row and second column of the MCORRELS output range B19:G24, and so the value of cell C19 is the correlation coefficient between Fund 1 and Fund 2 that we already computed in Figure 5.

Looking at the correlation array in B19:G24 of Figure 7, you can see that the lowest correlation coefficient is 0.5457, between Fund 2 (the transportation stocks) and Fund 3 (the utility stocks). The highest correlation coefficient between distinct funds is 0.9865, between Fund 1 and Fund 6, both of which are mixes of industrial stocks. Cells along the diagonal of the correlation array have the value 1 because the sample correlation of any data column with itself is always 1. The correlation array is always symmetric around this diagonal of 1s, because $CORREL(\text{column1}, \text{column2}) = CORREL(\text{column2}, \text{column1})$ for any two columns of data.

The range B32:G32 contains six random variables that have a Multivariate Normal joint probability distribution with the same means, standard deviations, and correlations that we have found for the annual growth ratios for these six funds. To make sure random variables that have the correlations in B19:G24, the array formula

$$\{=CORAND(B19:G24)\}$$

has been entered into the range B30:G30. Then these correlated random values have been used as inputs to NORMINV, along with the means from row 14 and the standard deviations from row 15, by entering the formula

$$=NORMINV(B30,B14,B15)$$

into cell B32, and then copying cell B32 to B32:G32.

	A	B	C	D	E	F	G	H
1	Annual growth ratios for six stock funds							
2		Fund 1	Fund 2	Fund 3	Fund 4	Fund 5	Fund 6	
3	1980	1.22	1.28	0.98	1.05	1.06	1.15	
4	1981	1.09	1.20	1.04	1.14	1.05	1.08	
5	1982	0.92	0.83	1.02	0.98	0.95	0.93	
6	1983	1.37	1.48	1.18	1.32	1.35	1.34	
7	1984	1.01	0.96	0.99	0.94	0.99	1.00	
8	1985	1.15	1.22	1.22	1.28	1.13	1.16	
9	1986	1.26	1.15	1.26	1.29	1.35	1.26	
10	1987	1.25	1.17	1.04	1.00	1.27	1.21	
11	1988	0.93	0.96	0.97	0.87	0.91	0.93	
12	1989	1.19	1.31	1.22	1.19	1.22	1.21	
13								
14	Means	1.138	1.155	1.092	1.106	1.126	1.129	
15	StDevs	0.152	0.192	0.115	0.162	0.162	0.140	
16								
17	Correlation coefficients							
18		Fund 1	Fund 2	Fund 3	Fund 4	Fund 5	Fund 6	
19	Fund 1	1.0000	0.8937	0.5980	0.7280	0.9178	0.9846	
20	Fund 2	0.8937	1.0000	0.5389	0.7404	0.7216	0.8804	
21	Fund 3	0.5980	0.5389	1.0000	0.8992	0.7627	0.7260	
22	Fund 4	0.7280	0.7404	0.8992	1.0000	0.7597	0.8109	
23	Fund 5	0.9178	0.7216	0.7627	0.7597	1.0000	0.9550	
24	Fund 6	0.9846	0.8804	0.7260	0.8109	0.9550	1.0000	
25								
26	A bad simulation model: independent Normals							
27		0.89	1.34	1.37	0.97	0.93	1.15	
28								
29	A good simulation model:							
30	CORANDs	0.2614	0.1034	0.1995	0.0851	0.3835	0.2433	
31	Multivariate Normals with appropriate correlations							
32		1.04	0.91	1.00	0.88	1.08	1.03	
33								
34	Portfolio \$ invested now in each fund							
35		3000	3000	1000	1000	1000	1000	
36					Total \$ now		10000	
37	Simulated portfolio returns next year (W)					E(W)	11328.90	
38		9844.42				Stdev(W)	1490.50	
39					10000	P(W<E39)	0.1863	
40	FORMULAS					VaR(5%)	8877.25	
41	B14.	=AVERAGE(B3:B12)		B14 copied to B14:G14.				
42	B15.	=STDEV(B3:B12)		B15 copied to B15:G15.				
43	B19:G24.	{=MCORRELS(B3:G12)}						
44	B27.	=NORMINV(RAND(),B14,B15)			B27 copied to B27:G27			
45	B30:G30.	{=CORAND(B19:G24)}						
46	B32.	=NORMINV(B30,B14,B15)			B32 copied to B32:G32.			
47	G36.	=SUM(B35:G35)						
48	G37.	=SUMPRODUCT(B35:G35,B14:G14)						
49	B38.	=SUMPRODUCT(B35:G35,B32:G32)						
50	G38.	=SUMPRODUCT(PRODS(B35:G35),PRODS(B15:G15),B19:G24)^0.5						
51	G39.	=NORMSDIST((E39-G37)/G38)			G40. =NORMINV(0.05,G37,G38)			

Figure 7. Building a simulation model of the six funds' annual growth ratios.

With these six cells B32:G32 simulating the performance of the six funds, we can make simulation models of any portfolio of investments in these funds. For example, suppose that we have a total of \$10,000 to invest, and we decide to put \$3000 each into Funds 1 and 2, and \$1000 each into Funds 3, 4, 5, and 6 now. With these investment levels listed in cells B35:G35, the value of this portfolio next year (\mathbf{W}) is simulated in cell B37 of Figure 7 by the formula

$$=\text{SUMPRODUCT}(B35:G35,B32:G32)$$

The expectation of this portfolio's value $E(\mathbf{W})$ is computed in cell G37 by the formula

$$=\text{SUMPRODUCT}(B35:G35,B14:G14)$$

as an application of our equation (3) for the expected value of a linear function. The standard deviation of this portfolio's value $\text{Stdev}(\mathbf{W})$ is computed in cell G38 by the formula

$$=\text{SUMPRODUCT}(\text{PRODS}(B35:G35),\text{PRODS}(B15:G15),B19:G24)^{0.5}$$

as an application of our equation (5) for the standard deviation of a linear function. That is, the standard deviation is the square root of the sum of the correlations (in B19:G24) multiplied by the corresponding pair-products of standard deviations (in B15:G15) and by the corresponding pair-products of investment levels (in B35:G35). This formula is quite long, but it is important not to forget the square root ($^{0.5}$) at the end. If we forgot the " $^{0.5}$ " at the end, then we would be computing $\text{Var}(\mathbf{W})$ instead of $\text{Stdev}(\mathbf{W})$!

Because the values of the assets in this portfolio are assumed to be Multivariate-Normal random variables, the total value of the portfolio \mathbf{W} is also a Normal random variable. So having its mean in cell G37 (\$11329) and its standard deviation in cell G38 (\$1490), we can compute the cumulative probability of \mathbf{W} being less than any value in cell E39, by the formula

$$=\text{NORMSDIST}((E39-G37)/G38)$$

as shown in cell G39 in Figure 7. With the value 10000 in cell E39, for example, we find that the probability of this portfolio next year being worth less than the \$10,000 initial investment is $P(\mathbf{W} < 10000) = 0.186$. The value at risk (VaR) at the 5% cumulative probability level for this portfolio can be computed by the formula

$$=\text{NORMINV}(0.05,G37,G38)$$

as shown in cell G40. Also, the formula $\text{NORMINV}(\text{RAND}(),G37,G38)$ would return a random variable with the same probability distribution as our simulated portfolio in cell B37.

A most common mistake in probability modeling is to make random variables independent where the evidence suggests that they should be strongly correlated. In spreadsheet models, this mistake can occur simply because it is easier to make random variables independent than to appropriately correlate them. For example, consider range B27:G27 of Figure 7, which contain a row of random variables that are made by entering

=NORMINV(RAND(),B14,B15)

into cell B27 and then copying cell B27 to B27:G27. That is, cells B27:G27 differ from the simulated growth ratios in cells B32:G32 only in that independent RAND() formulas are used in place of the CORAND values. So each cell in B27:G27 has exactly the same probability distribution as the corresponding cell in B32:G32, but cells B27:G27 are independent random variables, as they each depend on a different RAND. If you simulated the value of the portfolio of investments in B35:G35 with these independent growth ratios in B27:G27, using SUMPRODUCT(B35:G35,B27:G27) to simulate the portfolio value **W**, then you would get the same expectation as we found in cell G37, but you would find a much lower standard deviation than we computed in cell G38. (In fact, an application of equation (5i) above would yield the standard deviation 791 for this portfolio with independent growth ratios.) Any one simulated growth ratio in B27:G27 is as likely to be below 0.95 or above 1.30 (say) as the corresponding simulated growth ratio in B32:G32. But when one growth ratio takes a low value below 0.95 in B27:G27, there is still a substantial probability that some other independent random variable in B27:G27 will be take a high value above 1.30, so that the low return from one fund will be counterbalanced by a high return from the other. But our highly correlated empirical data suggests that, when one fund does badly, the others funds are also unlikely to do well, and so we really have a much greater probability that all of our investments may go badly together. Thus, cells B27:G27 must be considered inadequate and seriously misleading as a model of the annual growth ratios of these six stock funds.

6. Excel Solver and efficient portfolio design

We can now apply our sophisticated model of the stock market to a question of portfolio design. Suppose that we have been hired as consultants to advise an investor who has \$10,000

for the coming year. We might advise her to invest in the six funds that we have modeled in the previous section, as shown above in Figure 7. Indeed, we might well recommend that she use the investment plan shown in cells B35:G35 of Figure 7, where the investments now in Funds 1 through 6 are, respectively

3000, 3000, 1000, 1000, 1000, 1000

and the resulting portfolio value next year has expected value and standard deviation

$E = \$11,329$ and $Stdev = \$1490$

as shown in cells G37 and G38. We might point out the advantages of this diversified portfolio by noting that it offers an expected rate of return better than 13%, and that putting all \$10,000 in either of the two individual funds that offer better expected rates of return would be more risky. (Putting all \$10,000 in Fund 1 would yield standard deviation \$1520. Putting all \$10,000 in Fund 2 would yield standard deviation \$1922.) We are assuming here that this investor is risk averse, and does not simply want to maximize her expected portfolio value regardless of risk. That is, she might not prefer a portfolio that had a higher expected value if it also involved substantially more risk, as measured by the standard deviation.

But now our investor might well ask us whether there are other portfolios that yield lower standard deviations than this portfolio without decreasing the expected returns, and if so, which of these portfolios would be the least risky. With this question, we are facing an analytical problem that is quite different from what we have done up to now in our spreadsheets. If our client had asked about any single alternative portfolio, we could have easily evaluated it by entering the new investment levels into cells B35:G35 and reading the expected value and standard deviation of next year's returns in cells G37 and G38. But she is asking us to search among all possible ways of investing her \$10000, and find the one that has lowest standard deviation, subject to the constraint that it must offer expected return not lower than \$11,329. Such problems are called optimization problems.

Microsoft Excel comes with an add-in called Solver.xla which can solve optimization problems in spreadsheets. The default installation of Excel often does not install this xla file, and so you may need to redo a custom installation to get Solver.xla into the appropriate library folder in your computer's hard disk. Once it is there, you need to use the Tools>AddIns menu to

activate Solver.xla as a regular part of Excel. Once installed and added in, Solver should appear as a part of your Tools menu in Excel.

In the spreadsheet from Figure 7, let us now apply Solver to answer our investor's question about the least risky portfolio with the given expected return. With this spreadsheet activated, we launch Solver by the menu command sequence Tools>Solver, which opens the "Solver-Parameters" dialogue box. This dialogue box is looks complicated, but it basically requires us to specify three things.

First, we must specify a goal or target that we are directing Solver to pursue, by specifying in the Solver-Parameters box a target cell which is either to be maximized, or to be minimized, or to be made equal to some value. In this case, we have been asked to look for a portfolio that has as little risk as possible, and so we should select as our target cell G38, where the standard deviation of the portfolio's value is computed. We want to make this standard deviation to be as small as possible, and so we should select the Minimize option for this target cell.

Next, we must specify the cells that Solver is allowed to change or adjust in pursuit of this minimization target. We do so by clicking on the changing cells line (just above the middle of the Solver-Parameters box), and selecting there (by typing an address or by pointing with the mouse) the range of cells which Solver is allowed to change or adjust. Each of these cells must currently contain a simple numerical value, not a formula. In this case, the numbers that we want we want Solver to consider changing are the planned investments in the six stock funds, which have been entered into cells B35:G35 in Figure 7. So we should specify the range B35:G35 as the changing cells that we want Solver to adjust.

Finally, we must specify what constraints, if any, should restrict the values that Solver is allowed to use in the changing cells. In these constraints, various cells can be required to be less than some values, or equal to some values, or greater than some values. Specifying the constraints is often the hardest part in using Solver, because you can get nonsensical results if you misspecify your constraints. In this case, our mission is to invest \$10,000 in these six funds, and we are not interested in any investment plan that allocates more or less money. So we should require Solver, as it changes B35:G35, to keep the sum of the six investments equal to \$10,000. This sum of B35:G35 is computed by the formula in cell G36 in Figure 7, and so we should give

Solver the constraint

$$G36 = 10000$$

To enter this Solver constraint, click the "Add.." button in the Solver Parameters dialogue box, which launches the "Add-Constraint" dialogue box. Then enter the address G36 in the left-hand side of this Add-Constraint dialogue box, select the "=" option in the drop-down option list in the center of the Add-Constraint box, and enter the number 10000 in the right-hand side of this Add-Constraint box. Then click OK to close the Add-Constraint box and return to the Solver-Parameters box.

In the story told above, we had been asked to find the least risky portfolio that has an expected return not less than the \$11329 which was offered by the portfolio shown in Figure 7. The expected value returned by any changed B35:G35 portfolio would be computed in cell G37 in Figure 7. So we should also give Solver the constraint

$$G37 \geq 11329$$

To do so in the Solver-Parameters dialogue box, click the Add button to launch the Add-Constraint dialogue box again, and then enter the address G37 in the left-hand side of the Add-Constraint box, select the ">=" option in the center of the Add-Constraint box, and enter the number 11329 in the right-hand side of the Add-Constraint box. Then click OK to return again to the Solver-Parameters box.

So now we have specified the following Solver Parameters:

minimize the target cell G38 by changing cells B35:G35
subject to the constraints G36=10000, G37>=11329

Now we can click the "Solve" button in the Solver-Parameters box, and Solver will get to work, searching for new ways to invest \$10,000 that can yield lower risk without decreasing our expected value.

After some delay, the Solver-Results dialogue box should announce, "Solver has found a solution. All constraints and optimality are satisfied." This is good news, that Solver thinks that it has done its best to solve the problem that we gave it. Then with "Keep Solver Solution" checked, we should click OK to see Solver's solution. Unfortunately, in this case the solution may be one that our client will not appreciate, because the investments in Funds 1 through 6 that Solver

has recommended in cells B35:G35 are respectively

-17512, -1256, 2313, -10295, -18211, 54961

The negative numbers in this solution might be interpreted as instructions to sell short Funds 1, 2, 4, and 5. In simpler terms, Solver is suggesting that we should make a net investment of \$54,961 in Fund 6 and \$2,313 in Fund 3, which adds up to \$57,274, of which \$10,000 will come from the money that we have to invest, and the extra \$47,274 will be raised by selling shares in Fund 1 that are worth \$17,512, shares in Fund 2 that are worth \$1256, shares in Fund 4 that are worth \$10295, and shares in Fund 5 that are worth \$18,211. If our investor does not have these shares to sell, then perhaps some friend might let her borrow them, in exchange for a written promise to return them next year, which she could do by buying back the borrowed shares out of the returns from the sale of her positive investments. If our investor does not have a friend willing to make such a nice loan, then we must admit that a solution with negative values in cells B35:G35 is not really feasible for our investor.

To prevent such negative values in cells B35:G35, we should launch Solver again, go back to the Add-Constraint dialogue box, and enter as our third constraint

B35:G35>=0

(Solver will accept a constraint with a range of cells on the left-hand side and a single number on the right-hand side, in which case each cell's value is required to be above or below the given number.) Now when we click Solver's Solve button, it returns the following solution in cells B35:G35

6902, 1518, 1580, 0, 0, 0

This solution tells us that we can reduce our standard deviation to a minimum of \$1432 with nonnegative investments totaling \$10000, and with expected return not less than \$11,329, by investing \$6902 in Fund 1, \$1518 in Fund 2, and \$1580 in Fund 3.

This is just one example to illustrate the remarkable power of Solver to search for ways to improve some target cell by changing other cells, subject to well-specified constraints. You should consider other variations on this optimization problem, with other targets that may be of interest, and differently specified constraints.

For example, suppose that our investor becomes a bit more greedy and asks for the least

risky portfolio that would offer her an expected return (in cell G37) of at least \$12,000 next year (instead of just \$11,329). With the spreadsheet in Figure 7 still activated, we can launch Solver again, select the $G37 \geq 11329$ constraint, click the "Change" button, and change the constraint to $G37 \geq 12000$.

But now when we ask Solver to solve this new optimization problem, it will announce in its Solver-Results box that "Solver could not find a feasible solution." That is, our optimization problem is infeasible, which means that our constraints are impossible to satisfy and must be relaxed in some way. With nonnegative investments that sum to \$10,000, we cannot get a higher expected return than \$11,548, which is achieved by putting all \$10,000 into Fund 2. To verify this fact, you should launch Solver again and give it the optimization problem:

maximize the target G37 (the expected return) by changing cells B35:G35
subject to the constraints $G36 = 10000$, $B35:G35 \geq 0$

(Note: you can eliminate a selected constraint by the "Delete" button in the Solver-Parameters dialogue box.) For this new problem, Solver should return the solution

0, 10000, 0, 0, 0, 0

in cells B35:G35.

To see one of the more puzzling outcomes that you can get from running Solver, try the following optimization problem in the spreadsheet from Figure 7:

maximize the target cell G37 by changing cells B35:G35
subject to the constraint $G36 = 10000$.

That is, we are asking Solver to find the highest possible expected return for a portfolio that costs \$10,000 now, but now we are allowing some funds to have negative investments (selling them short) because the nonnegativity constraint has been deleted. When you ask Solver to solve this problem, the Solver-Results box will report "The Set Target Cell values do not converge." This obscure message actually means that Solver has found an unbounded optimal solution. That is, Solver has found a way to go on improving your target cell infinitely. In this case, Solver is telling us that it can find ways to earn expected returns in the millions or billion of dollars next year, with a net investment of only \$10,000 now. So whenever you get this "Set Target Cell values do not converge" message, it means either that Solver has found a way for you to make an

infinite amount of money, or else (as is more likely) you have misspecified the problem. In this case, the problem is that we are ignoring risk and credit constraints, by assuming that we can raise millions of dollars now by short-selling the low-expected-return funds (such as Fund 3), and then invest these millions into the high-expected-return funds (such as Fund 2). That is, some constraint that rules out enormous risks or negative investment levels needs to be specified our optimization model, if it is to describe the investor's real situation.

Now I have to tell you one more thing about Solver, which may explain why Microsoft omits it from the default installation of Excel: Unlike most other features in Excel, Solver can often give the wrong answer. To illustrate the problem consider the simple spreadsheet shown in Figure 8. In a blank spreadsheet, let us start by entering any positive value (say the value 1) into cell A1. Then enter the formula =A1^2 into cell B1. Now launch Solver, and give it the optimization problem:

Maximize the target cell B1 by changing cell A1
subject to the constraints $A1 \leq 5$, $A1 \geq -10$.

Solver will find and report the solution that is as shown in Figure 8, where the value of cell A1 has been changed to 5, yielding the target-cell value of 25. But we could get a much higher target-cell value of 100 if we entered the value -10 into cell A1, which also satisfies the required constraints. Why did Solver miss this obviously better solution?

	A	B	C	D
1	5	25		
2				
3	FORMULAS FROM RANGE A1:B1			
4	B1.	=A1^2		
5				
6	Solver: Maximize B1 by changing A1			
7	subject to $A1 \leq 5$, $A1 \geq -10$.			

Figure 8. A simple example where Solver can fail.

To understand what went wrong, you need to know something about how Solver works. Whenever we launch Solver, it starts with some numerical values already entered (by us) into the changing cells. Suppose first that our initial values do not violate any constraints. Then Solver considers a variety of very small changes in these changing-cell values, looking for ways to make

small changes that will improve the target-cell value without violating any constraints. If it finds any way to do this, then it continues making larger changes in the same direction for as far as it can get further improvement in the target cell without violating any constraints. Then Solver repeats this process, looking around the new changing-cell values to find a direction in which small changes can improve the target-cell value without violating any constraints, and it again moves the changing-cell values in this new direction as long as further improvement is possible. When Solver finds changing-cell values from which there are no small changes that could offer improvement without violating constraints, then it stops and reports that an optimum has been found. Such a solution where small feasible changes cannot improve the target is called a local optimal solution or a local optimum. (If our initial changing-cell values do not satisfy the constraints, then Solver works similarly to minimize the violation of the constraints, turning to target-cell improvement after all constraint violations have been eliminated.)

For the example in Figure 8, Solver has found a local optimum at $A1=5$, where small increases in $A1=5$ would not be allowed and small decreases below $A1=5$ would not improve the target. But this local optimum is not really the optimal solution because, decreasing $A1$ by a larger amount (subtracting 10 or more) could increase the target cell without violating any constraints. And indeed, if we had started with any negative value in $A1$ when we launched Solver, then it would have moved swiftly to the true optimal solution at $A1=-10$.

You can think of Solver as like a robot who searches for the highest point in the world by continually walking in an uphill direction, until it reaches the top of a hill or mountain, where it stops. Such a robot is not likely to find the top of Mt. Everest unless it starts in the right part of Nepal or Tibet! Some optimization problems may resemble a smooth landscape dominated by one large hill, and Solver will be reliably successful in such cases. But other optimization problems may resemble a rough landscape with many separate mountain ranges, and the results reported by Solver may be very different, depending critically on its initial position.

So whenever you use Solver to analyze an optimization problem, you should run Solver several times, with a variety of different initial values for the changing cells. If Solver always comes back reporting the same optimal solution, then you may have some confidence that it is truly optimal. But if Solver's optimal solution changes each time, then you should run it many

times and keep a record of the changing-cell and target-cell values for the best solutions that it reports. Unless the optimization problem has some special mathematical structure, however, you might never be sure that one more application of Solver would not find an even better solution.

7. Using CORAND with nonNormal random variables

We have learned to make many type of random variable (Normal, Lognormal, Triangular, Gamma, etc.) by using RANDs as the first parameter in the appropriate inverse-cumulative function (NORMINV, LNORMINV, TRIANINV, GAMINV, etc.). In each case, if we use a CORAND value in place of the RAND value, we will get a random variable drawn from the same probability distribution, because each CORAND has the same uniform distribution as the RAND. But with CORAND, these random variables can be positively or negatively correlated with other random variables that are made using other random values in the same CORAND array. So CORAND can be used as a versatile tool for creating nonindependent random variables with different kinds of probability distributions.

But when CORAND values are used in this way to make random variables, the correlation of the resulting random variables can be guaranteed to equal the correlation parameter in the CORAND function only when we are making Normal random variables. For nonNormal random variables, the correlation of the resulting random variables may be different from the correlation parameter of the CORANDs. The CORANDs themselves have a correlation that differs only slightly from the value of their input parameter. But when a pair of positively correlated CORANDs are used to make two random variables that are skewed in opposite directions, these random variables often have a correlation that is much smaller than the CORANDs' parameter.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Correlation parameter				Quartile parameters						FORMULAS			
2	0.9			Q1	-2	-3				B9:C9. {=CORAND(A2)}				
3				Q2	0	0				SimTable data fills B10:C110.				
4				Q3	3	2				C7. =CORREL(B10:B110,C10:C110)				
5								Normalized rank		E9. =GENLINV(B9,E\$2,E\$3,E\$4)				
6	CORREL(Corand data)			CORREL(GenLogNmls)			correlation			F9. =GENLINV(C9,F\$2,F\$3,F\$4)				
7			0.8963			0.6963			0.8956	E9:F9 copied to E10:E110				
8	Table of 101 corands			GenLognormals						F7. =CORREL(E10:E110,F10:F110)				
9	SimTabl	0.4726	0.6007		-0.243	0.853	Normalized data			H10:H110. {=NORMIZE(E10:E110)}				
10	0	0.1268	0.2475		-2.98	-3.042		-4.648	-4.132	I10:I110. {=NORMIZE(F10:F110)}				
11	0.01	0.1437	0.1038		-2.834	-6.799		-4.355	-6.452	I7. =CORREL(H10:H110,I10:I110)				
12	0.02	0.8658	0.6683		5.6713	1.3811		5.5921	0.3154					
13	0.03	0.4989	0.715		-0.01	1.7359		1.1099	1.0068					
14	0.04	0.8881	0.88		6.4674	3.0389		6.1081	3.6252					
15	0.05	0.8126	0.8124		4.2297	2.4792		4.5601	2.128					
16	0.06	0.0676	0.0154		-3.556	-15.97		-6.269	-12.09					
17	0.07	0.1999	0.3416		-2.383	-1.669		-2.624	-2.892					
18	0.08	0.1847	0.2779		-2.502	-2.55		-2.99	-3.693					
19	0.09	0.3566	0.2865		-1.189	-2.419		-0.734	-3.416					
20														
21														
22														
23														
24														
25														
26														
27														
28														
29														
30														
31														
32														
33														
34														

Figure 9. NonNormal random variables and normalized rank correlations.

To illustrate this effect, cells B9:C9 in Figure 9 contain the array formula

`{=CORAND(A2)}`

and cell A2 contains the value 0.9. So two Normal random variables that are driven by these two CORANDs would have correlation 0.9. But cells E9 and F9 contain nonNormal random variables driven by these two CORANDs. The formula in cells E9 and F9 respectively are

`=GENLINV(B9,E$2,E$3,E$4)` and `=GENLINV(C9,F$2,F$3,F$4)`

The values in (E2,E3,E4) are (-2,0,3), and the values in (F2,F3,F4) are (-3,0,2). So these two Generalized-Lognormal random variables are not Normal (because their quartiles do not satisfy the equal-differences property). Cell E9 is positively skewed, with a long tail on the high side. Cell F9 is negatively skewed, with a long tail on the low side.

Cells B10:C110 in Figure 9 contain data from 101 recalculations of the CORANDs in cells B9:C9. That is, each row in B10:C110 contains a pair of CORAND values with the correlation parameter 0.9, and the values in each row are independent of the values in every other row. The GENLINV formulas from cells E9:F9 have been to E10:F110, and so the E and F cells in each row from 10 to 110 contain a pair of Generalized-Lognormal random variables made with 0.9-correlated CORANDs. But the formula

`=CORREL(E10:E110,F10:F110)`

in cell F7 returns the value 0.6963. This value is just a statistical estimate from 101 sampled pairs, but it is actually an accurate estimate of the true correlation of these two random variables.

The scatter charts at the bottom of Figure 9 show why this happens. When two random variables have a high correlation close to 1, it means that the values of these random variables tend to be scattered close to a positively sloped line when they are plotted together in an XY-scatter chart. But the different skews of the random variables in the E and F columns introduces a nonlinearity which quite evident in the middle scatter plot. Because of the high positive correlation parameter in the underlying CORANDs, the E and F values (which both have median 0) tend to be either both greater than 0 or both less than 0. But the E value tends to go much farther from 0 than the F value when they are positive, while the F value tends to go much farther from 0 than the E value when they are negative. As a result, the (E,F) pairs in the middle chart are scattered along an obtuse angle. Because these points do not stay close to any one

straight line, their correlation is much less than 1.

But now we have a tricky problem: Suppose that we have data for two continuous random variables that are clearly not independent and not Normal, and we want to simulate these two random variables using CORAND. What parameter for CORAND should we use to get our simulated output to be correlated like our given data? So Figure 9 shows that, to get Generalized-Lognormals with quartile points $(-3,0,2)$ and $(-2,0,3)$ to have a statistical correlation of about 0.7, we may need CORANDs' input parameter to be around 0.9. But, if we cannot estimate the appropriate parameter for CORAND by applying the CORREL function to our data for nonNormal random variables, then how can we estimate it? The answer is that, for any two continuous random variables, the appropriate parameter for CORAND can be estimated by applying the CORREL function, not to our original data for these two random variables, but to order-preserving normalizations of these two data sets that are returned by the NORMIZE array function in Simtools.

Simtools's NORMIZE array function is illustrated in columns H and I of Figure 9. The range H10:H110 was selected and filled with the array formula

$$\{=NORMIZE(E10:E110)\}$$

This array formula fills H10:H110 with numerical values that are ordered the same way as the E10:E110. (For example, the pattern of inequalities $E10 < E11 < E12 > E13$ is repeated by the corresponding normalized values $H10 < H11 < H12 > H13$.) Also, the mean and standard deviation of these values in H10:H110 are almost the same as the mean and standard deviation of E10:E110. But the numerical values in H10:H110 have been distributed by NORMIZE to fit a Normal distribution.

Next, by copying H10:H110 to I10, the range I10:I110 in Figure 9 has been filled with the array formula

$$\{=NORMIZE(F10:F110)\}$$

which returns a similar normalization of the data in F10:F110. The statistical correlation of these two normalized data arrays in the H and I columns is computed in cell I7 by the formula

$$=CORREL(H10:H110,I10:I110)$$

This correlation, which may be called the normalized rank correlation of the original data in the E

and F columns, is our recommended estimate of the parameter for CORAND that could generate such data. Notice that the value of this normalized rank correlation in cell I7 (0.8956) is indeed quite close to the original CORAND parameter in A2 (0.9). More generally, when we use a CORAND array to drive a pair of nonNormal continuous random variables, we may refer to the correlation parameter of the CORANDs as the normalized rank correlation of the resulting nonNormal random variables, because it could be estimated from data in this way.

For completeness, let me tell you how you could compute the NORMIZE values in cells H10:H110 of Figure 9 without Simtools. First, enter the formula =AVERAGE(E10:E110) in cell P1, enter the formula =STDEV(E10:E11) in cell P2, and enter the formula =COUNT(E10:E110) in cell P3. Next, in cell P10 enter the formula

$$=NORMINV((RANK(E10,E10:E110,1)-0.5)/P3, P1, P2)$$

Then copy P10 to P10:P110. Now suppose that the average of E10:E110 in P1 is 1.66, the standard deviation in P2 is 4.36, and the count in P3 is 101. Then corresponding to the smallest entry in E10:E110, we will find the value NORMINV(0.5/101,1.66,4.36) where its row meets column P. Corresponding to the second-smallest entry in E10:E110, we will find the value NORMINV(1.5/101,1.66,4.36) where its row meets column P. And so on, up to the largest entry in E10:E110 which has the corresponding value NORMINV(100.5/101,1.66,4.36) in column P. These results in P10:P110 should match the NORMIZE values in H10:H110.

The NORMIZE array function is awkward and slow, and it can interact badly with calculation bugs in some versions of Office 97. So for many random variables that are even approximately Normal, it may be best simply to use the statistical correlations of the our (untransformed) data to estimate the appropriate CORAND parameters. But for very skewed random variables, it may be worthwhile also to look at the normalized rank correlation.

8. Subjective assessment of correlations

In the preceding chapter, we emphasized subjective quartile assessment, as a practical way of assessing probability distributions for unknown quantities even when relevant data for statistical analysis is not available. But quartile boundary points, like means and standard deviations, only measure beliefs about one unknown quantity at a time. It is just as important to

measure beliefs about the relationships between unknown quantities, such as are expressed by their pairwise correlations. So in this section, we develop a practical method for subjectively assessing correlations (or, more accurately, normalized rank correlations).

To begin, let \mathbf{X} and \mathbf{Y} denote two unknown quantities that we perceive as not independent. Suppose that we have subjectively assessed quartile boundary points for each of these unknown quantities. For example, \mathbf{X} may denote the sales (in 1000s of units) of some proposed new product in the first year after it is introduced, and \mathbf{Y} may denote the average annual sales (again in 1000s) of the same product over its first ten years. Suppose that the responsible marketing manager has assessed for the first-year sales \mathbf{X} four equally likely quartiles with boundary points 15, 33, and 75, and has assessed for the ten-year average sales \mathbf{Y} equally likely quartiles with boundary points 40, 80, and 160. These subjectively assessed quartile boundary points are entered in cells B5:C7 of the spreadsheet shown in Figure 10.

Now suppose that our plan is to simulate \mathbf{X} and \mathbf{Y} by Generalized-Lognormal random variables that are driven (in the first parameter for their GENLINV formulas) by a pair of linked CORAND values, as illustrated in cells B21:C21 of Figure 10. So we need to ask our expert some question that can be used to estimate the correlation parameter for these CORANDs.

To assess the relationship between these two unknowns, we should ask some question about how beliefs about one unknown would be affected by getting information about the other. If we perceived these unknown quantities to be independent then learning about one would have no effect on our beliefs about the other, but if they are highly correlated then learning about one would greatly change our beliefs about the other.

One such question that we may ask our expert is:

"If you learned that the quantity \mathbf{X} was actually equal to the value that you have assessed for \mathbf{X} 's upper quartile boundary point but you learned nothing else relevant to the unknown \mathbf{Y} , then for what number \hat{y} would you say that \mathbf{Y} was equally likely to be above or below \hat{y} ?"

That is, we let x_3 denote this upper quartile boundary point for \mathbf{X} , then we are asking for a number \hat{y} such that

$$P(\mathbf{Y} < \hat{y} | \mathbf{X} = x_3) = 1/2.$$

This number \hat{y} can be called the conditional median of **Y** given that **X** is equal to x_3 .

	A	B	C	D	E	F	G
1	WORKSHEET FOR SUBJECTIVE ASSESSMENT OF CORRELATIONS						
2	Suppose X and Y have Generalized-Lognormal distributions						
3	with quartile boundary points as listed below:						
4		X	Y				
5	Q1 (0.25)	15	40				
6	Q2 (0.50)	33	80				
7	Q3 (0.75)	75	180				
8							
9	When Correlation(X,Y) =			0.8			
10	0.2947391	= Conditional median of one CORAND given other is 0.25					
11	0.7052609	= Conditional median of one CORAND given other is 0.75					
12		X					
13		17.492973	= Conditional median of X given Y=40				
14		63.542873	= Conditional median of X given Y=180				
15			Y				
16			45.363323	= Conditional median of Y given X=15			
17			152.09218	= Conditional median of Y given X=75			
18	Simulation model:						
19	(corands)	0.3567044	0.4486018				
20		X	Y				
21		21.357971	69.268578				
22							
23	FORMULAS						
24	A10.	=NORMSDIST(D9*NORMSINV(0.25))					
25	A11.	=NORMSDIST(D9*NORMSINV(0.75))					
26	B13.	=GENLINV(A10,B5,B6,B7)		C16.	=GENLINV(A10,C5,C6,C7)		
27	B14.	=GENLINV(A11,B5,B6,B7)		C17.	=GENLINV(A11,C5,C6,C7)		
28	B19:C19.	{=CORAND(D9)}					
29	B21.	=GENLINV(B19,B5,B6,B7)		C21.	=GENLINV(C19,C5,C6,C7)		
30	C13.	=" = Conditional median of X given Y="&C5					
31	C14.	=" = Conditional median of X given Y="&C7					
32	D16.	=" = Conditional median of Y given X="&B5					
33	D17.	=" = Conditional median of Y given X="&B7					

Figure 10. Computing conditional medians for subjective correlation assessment.

In our example, we might put this question to the marketing manager as follows. "You have just told us that you think ten-year average sales are equally likely to be above or below 80. But now suppose for a moment that first-year sales turned out to be 75. Given this information, you might want to revise your beliefs about the ten-year average of sales that we will achieve. So if you knew that first-year sales were 75, for what number \hat{y} would you then be indifferent

between betting that the ten-year average sales will be greater than \hat{y} , or betting that the ten-year average sales will be less than \hat{y} ?" Suppose that, after some thought, the marketing manager reports 150 as her subjectively assessed conditional median of Y given $X=75$.

To compute the appropriate correlation parameter for our CORANDs from such a conditional median, you need to know one more formula. When a pair of CORAND values is generated with some correlation-parameter r , then for any number p between 0 and 1, the formula

$$\text{NORMSDIST}(r*\text{NORMSINV}(p))$$

gives us the conditional median for one of these CORAND values given that the other is equal to p (or is very close to p). (Note: $\text{NORMSINV}(p)$ in Excel is the same as $\text{NORMINV}(p,0,1)$.)

In Figure 10, for example, cells B19:C19 contain the array formula

$$\{=\text{CORAND}(D9)\}$$

and cell A11 contains the formula

$$=\text{NORMSDIST}(D9*\text{NORMSINV}(0.75))$$

When cell D9 is equal to 0.8, this formula returns the value 0.705 in cell A11. So if we made a table of thousands of simulations of the CORANDs in cells B19:C19 with the correlation parameter D9 equal to 0.8, and if we then selected from this simulation data the rows where the B cell had value very close to 0.75, then the value of the C cell should be less than cell A11 (0.705) in about half of these selected simulations. (To select 200 rows where the value of B19 was between 0.74 and 0.76, we would need about 10,000 total simulations.) But these CORANDs in B19:C19 are being used as the random first parameter driving the simulated values of X and Y in cells B21 and C21. So a B19 value near 0.75 implies a simulated value of X near $\text{GENLINV}(0.75,15,33,75) = 75$ (because 75 is the quartile boundary point with cumulative probability 0.75 for the unknown quantity X). And a C19 value less than the value in cell A11 implies a simulated value of Y less than

$$\text{GENLINV}(A11,40,80,180)$$

which is computed in cell C17 of Figure 10. When $D9 = 0.8$ and $A11 = 0.705$, this formula in cells C17 takes the value 152. So if the correlation parameter were 0.8, then having X near 75 would imply that Y was equally likely to be above or below 152, as is calculated in cell by the formula

But our expert subjectively assessed 150 as the conditional median of Y given $X=75$, and so the correlation 0.8 is not quite right. By adjusting the value of cell D9 in this spreadsheet and watching how cell C11 changes, you can verify that the correlation $D9 = 0.783$ makes the computed conditional median in cell C17 very close to the subjectively assessed value of 150. So a CORAND correlation parameter of 0.783 will give us simulated X and Y values that fit the expert's subjective assessment.

Cell C16 in Figure 10 similarly computes the conditional median of our simulated Y when the simulated X is very close to 15 (the quartile point with cumulative probability 0.25 for the unknown quantity X). When the correlation in cell D9 is 0.783, the value of cell C16 is about 46. So to check our correlation assessment, we might ask the expert "If I had told you first-year sales would be 15, might you say then that the average annual sales over ten years would be about as likely to be below 46 as above 46?" If the marketing expert felt comfortable with this statement, then we would have further support for estimating 0.783 to be the normalized rank correlation between first-year sales and ten-year average annual sales for this new product. Otherwise, we might try to modify the correlation in D9 to until the computed values in C16 and C17 both seem approximately correct to the expert as conditional medians of Y given the X values of 15 and 75 respectively.

9. Summary

This chapter covered the use of correlations to measure the relationship between random variables. Joint distributions of discrete random variables were introduced first. The covariance of two random variables was defined as the expected value of the product of their deviations from their respective means, and their correlation was defined as their covariance divided by the product of their standard deviations. Among these two measures of the relationship between random variables, we emphasized the correlation because it is a unit-free number between -1 and 1 , and it is somewhat easier to interpret. We saw correlation arrays for two or more random variables, which display the correlations among all pairs of these random variables, and always have 1's along the diagonal from top-left to bottom-right.

We studied some general formulas for linear functions of random variables. When the

means, standard deviations, and pairwise correlations for a set of random variables are known, we learned general formulas for computing the expected values, variances, standard deviations, and covariances of any linear functions of these random variables. We also saw the simpler special form of these formulas when the underlying random variables are independent.

In Section 4 we learned about Multivariate-Normal random variables. These random variables have a joint probability distribution that is parameterized by their individual means, individual standard deviations, and their pairwise correlations. It is very convenient to work with Multivariate-Normal random variables because any linear function of them has a Normal probability distribution. Multivariate-Normal random variables can be constructed by making linear functions that all depend on a common collection of underlying independent Normals. But we learned an easier way to construct Multivariate-Normal random variables with any given means, standard deviations, and correlations, by using the CORAND array function to provide appropriately correlated random inputs to NORMINV.

We learned how correlations are estimated from sample data by the CORREL function. We used MCORRELS to make correlation arrays from sample data for several random variables.

CORAND can be used to generate correlated random variables with other probability distributions, but the statistical correlation of nonNormal random variables may be different from the correlation parameter of the CORANDs that drive them. In general, the CORANDs' correlation parameter actually corresponds to a normalized rank correlation of the resulting random variables. The difference between this normalized rank correlation and the usual statistical correlation is often quite small, however, and it vanishes for Normal random variables.

Finally, we also introduced a technique for subjective assessment of the correlation between two unknown quantities, by asking about how the conditional median for one unknown quantity would depend on given information about the other unknown quantity.

Two important Excel techniques were introduced in this chapter: array formulas and Solver. Array formulas are entered into a range of cells with the special keystroke [Ctrl]-[Shift]-[Enter]. To use Solver, we learned how to specify the target cell to be maximized or minimized, the changing cells, and the constraints that make up an optimization problem. We learned to recognize solution difficulties such as infeasibility, unbounded optimality, and local

optimality. Because of the possibility of a local optimal solution that is not the true optimum, we saw the importance of rerunning Solver with different initial conditions.

The Excel function CORREL was introduced in this chapter. We also used several Simtools functions that have been designed to make correlation analysis easier: COVARPR, CORRELPR, and the array functions PRODS, CORAND, MCOLL, and NORMIZE.

PROBLEMS FOR CHAPTER 4.

1. X denote the dollar value per share that will be returned by Stock 1 next year, and let Y denote the dollar value per share that will be returned by Stock 2 next year.

In a simple discrete model, we figure that the possible values of X are \$40, \$45, and \$50, while the possible values of Y are \$8, \$9, and \$10, and the joint probability distribution for X and Y is

x	y	$P(X=x \ \& \ Y=y)$
40	8	0.15
40	9	0.05
40	10	0
45	8	0.10
45	9	0.25
45	10	0.10
50	8	0
50	9	0.15
50	10	0.20

- (a) Calculate the expected values $E(X)$ and $E(Y)$.
- (b) Calculate the standard deviations $Stdev(X)$ and $Stdev(Y)$
- (c) Calculate the covariance $Covar(X,Y)$ and the correlation $Correl(X,Y)$.
- (d) Suppose that Stock 1 is currently selling for \$42.50 per share and Stock 2 is currently selling for \$8.50 per share. An investor currently owns 600 shares of Stock 1 and has no shares of Stock 2, but she is thinking about selling some of her shares of Stock 1 and investing the proceeds in Stock 2 now. Make a table showing how the expected value and standard deviation of her portfolio next year would depend on how many shares of Stock 1 she keeps. In this table, just consider the possibility of selling whole blocks of 100 shares, so that she might keep 0 shares, 100 shares, 200 shares, etc., up to 600 shares of Stock 1.

2. Suppose that we are planning to make investments in three mutual funds. Suppose that the annual growth ratios of the three funds are Multivariate Normal random variables with the following means, standard deviations, and correlations.
 The annual growth ratio of Fund 1 has expected value 1.12 and standard deviation 0.15.
 The annual growth ratio of Fund 2 has expected value 1.14 and standard deviation 0.18.
 The annual growth ratio of Fund 3 has expected value 1.10 and standard deviation 0.19.
 The annual growth ratios of Funds 1 and 2 have correlation 0.8.
 The annual growth ratios of Funds 1 and 3 have correlation 0.5.
 The annual growth ratios of Funds 2 and 3 have correlation 0.3.
 We have a \$50,000 in cash that we can invest now. Any cash that we do not invest in these mutual funds will be put into risk-free short-term bonds that pay 5% annual interest (not compounded).

- (a) Set up a spreadsheet to compute, for any level of investments in these three funds:
- (i) the expectation of our portfolio's value one year from now,
 - (ii) the standard deviation of our portfolio's value one year from now,
 - (iii) the probability that our portfolio's value one year from now will be less than \$50,000,
 - (iv) and the 5%-cumulative value at risk of our portfolio's value one year from now.
- Compute these quantities when for the plan of investing \$20,000 in Fund 1, \$15,000 in Fund 2, \$10,000 in Fund 3, and \$5000 in the risk-free bonds.
- (b) Find the portfolio that minimizes the standard deviation of our portfolio's value next year, subject to the constraint that the expected return should not be less than \$55,000. (Assume that we cannot borrow money or sell-short any mutual funds.)
- (c) How does your answer to part (b) change if we add the constraint that all \$50,000 must be invested now in the mutual funds? (That is, nothing can be in bonds.)

SUPERIOR SEMICONDUCTOR -- Part C

Eastmann felt very pleased with her analysis of the decision to enter the T-regulator market. She had developed an analytical model that took account of all the uncertainties confronting by her colleagues at Superior Semiconductor. Using on this model, she could also offer a good quantitative summary of the potential benefits and risks of this project, and she was prepared to give a bottom-line recommendation in favor of the project based on its positive expected value.

But something bothered her as she looked again at her spreadsheet model. The easiest thing had been to assume that all the unknown quantities were independent. But was this really an appropriate assumption? In particular, there was good reason to think that the number of competitors entering the market would not be independent of market size. Given the currently available information, Suttcliff had been willing to say that the probability of 5 competitors entering was only about 0.10. But if he learned that the total discounted value of the market was on the large side, say larger than \$125 million, then he might reasonably increase his subjective probability of 5 competitors entering to 0.20 or higher.

So Eastmann decided to do sensitivity analysis on the correlation between the total market value and the number of entrants (which she had previously assumed to be zero, when she made them independent random variables in her model). She wanted to see how Superior Semiconductor's expected profit would change in her model if she changed the model by giving these two random variables a normalized rank correlation of 0.4, or 0.6, or 0.8.

If the correlation did affect the bottom-line results, then she would need some intuitive way to say which correlation was "correct." So she decided to also tabulate, as a function of this correlation coefficient, how the conditional probability of 5 competitors given a total market value over \$125 million would depend on the assumed correlation coefficient. Then she could ask Suttcliff to give her his subjective assessment of this conditional probability, and she could base her final recommendations on the model with correlation that best approximated his subjective assessment.