

Practical Probability with Spreadsheets
Chapter 2: DISCRETE RANDOM VARIABLES

Case: SUPERIOR SEMICONDUCTOR (Part A)

Peter Suttcliff, an executive vice-president at Superior Semiconductor, suspected that the time might be right for his firm to introduce the first integrated T-regulator device using new solid-state technology. This new product seemed the most promising of the several ideas that had been suggested by the head of Superior's Industrial Products division. So Suttcliff asked his staff assistant Julia Eastmann to work with Superior's business marketing director and the chief production engineer to develop an evaluation of the profit potential from this new product.

According to Eastmann's report, the chief engineer anticipated substantial fixed costs for engineering and equipment just to set up a production line for the new product. Once the production line was set up, however, a low variable cost per unit of output could be anticipated, regardless of whether the volume of output was low or high. Taking account of alternative technologies available to the potential customers, the marketing director expressed a clear sense of the likely selling price of the new product and the potential overall size of the market. But Superior had to anticipate that some of its competitors might respond in this area by launching similar products. To be specific in her report, Eastmann assumed that 3 other competitive firms would launch similar products, in which case Superior should expect 1/4 of the overall market.

Writing in the margins of Eastmann's report, Suttcliff summarized her analysis as follows:

- Superior's fixed set-up cost to enter the market: \$26 million
- Net present value of revenue minus variable costs in the whole market: \$100 million
- Superior's predicted market share, assuming 3 other firms enter: 1/4
- Result: predicted net loss for Superior: (\$1 million)

"Your estimates of costs and total market revenues are probably very accurate," Suttcliff told Eastmann. "But your assumption about the number of other firms entering to share this market with us is just a guess. I can count 5 other semiconductor firms that might seriously consider competing with us in this market. In the worst possible scenario, all 5 of these firms could enter the market, although that is rather unlikely. There is no way that we could keep this market to ourselves for any length of time, and so the best possible scenario is that only 1 other firm would enter the market, although that is also rather unlikely. I would agree with you that the most likely single event is that 3 other firms would enter to share the market with us, but that event is only a bit more likely than the possibilities of having 2 other firms enter, or having 4 other firms enter. So there is really a lot of uncertainty about this situation, and your analysis might be more convincing if you did not ignore it."

"We can redo the analysis in a way that takes account of the uncertainty by using a probabilistic model," Eastmann replied. "The critical step is to assess a probability distribution for the unknown number of competitors who would enter this market with us. So I should try to come up with a probability distribution that summarizes the beliefs that you expressed." Then after some thought, she wrote the following table and showed it to Suttcliff:

<u>K</u>	<u>Probability that K other competitors enter the market</u>
1	0.10
2	0.25
3	0.30
4	0.25
5	0.10

Suttcliff studied the table. "I guess that looks like what I was trying to say. I can see that your probabilities sum to 1, and you have assigned higher probabilities to the events that I said were more likely. But without any statistical data, is there any way to test whether these are really the right probability numbers to use?"

"In a situation like this, without data, we have to use subjective probabilities," Eastmann explained. "That means that we can only go to our best expert and ask him whether he believes each possible event to be as likely as our probabilities say. In this case, if we take you as best expert about the number of competitive entrants, then I could test this probability distribution by asking you questions about your preferences among some simple bets. For example, I could ask you which you would prefer among two hypothetical lotteries, where the first lottery would pay you a \$10,000 prize if exactly one other firm entered this market, while the second lottery would pay the same \$10,000 prize but with an objective 10% probability. Assuming that you had no further involvement with this project, you should be indifferent among these two hypothetical lotteries if your subjective probability of one other firm entering is 0.10, as my table says. If you said that you were not indifferent, then we would try increasing or decreasing the first probability in the table, depending on whether you said that the first or second lottery was preferable. Then we could test the other probabilities in the table by similar questions. But if we change any one probability in my table then at least one other probability must be changed, because the probabilities of all the possible values of the unknown quantity must add up to 1."

Suttcliff looked again at the table of probabilities for another minute or two, and then he indicated that it seemed to be a reasonable summary of his beliefs.

1. Introduction

Uncertainty about numbers is pervasive in all management decisions. How many units of a proposed new product will we sell in the year when it is introduced? How many yen will a dollar buy in currency markets a month from today? What will be the closing Dow Jones Industrial Average on the last trading day of this calendar year? Each of these number is an unknown quantity. If the profit or payoff from a proposed strategy depends on such unknown quantities, then we cannot compute this payoff without making some prediction of these unknown quantities.

The usual approach to this problem is to assess your best estimate for each of these unknown quantities, and use these estimates to compute the bottom-line payoff for each proposed strategy. Under this method of point-estimates, the optimal strategy is considered to be the one that gives you the highest payoff when all unknown quantities are equal to your best estimates.

But there is a serious problem with this method of point-estimates: it completely ignores your uncertainty. In this course, we will study ways to incorporate uncertainty into the analysis of decisions. Our basic method will be to assess probability distributions for unknown quantities, and then to create random variables to simulate these unknown quantities in spreadsheet simulation models. (Note on terminology: The term "random variable" could be taken by definition to mean the same thing as the phrase "unknown quantity." But as a matter of style here, we will tend to reserve the term "unknown quantity" for unknowns in the real world, and "random variable" will be used more for values in spreadsheets that are unknown because they depend on unknown RAND values.)

To illustrate these ideas, we consider the Superior Semiconductor case, Part A. In this case, we have a decision about whether our company should introduce a proposed new product. It is estimated that the fixed cost of introducing this new product will be \$26 million. The total value of the market (price minus variable unit costs, multiplied by total demand) is estimated to be \$100 million. It is also estimated that 3 other firms will enter this market and share it with us. Thus, by the method of point-estimates, we get a net profit (in \$millions) of $100/(3+1) - 26 = -1$, which suggests that this product should not be introduced. But all the quantities in this calculation (fixed cost, value of the market, number of competitive entrants) are really subject to some uncertainty. We will see, however, that when uncertainty is properly taken into account, the

new product may be recognized as worth introducing.

The analysis in Part A of this case focuses on just one of these unknowns: the number of entrants. Uncertainty about other quantities (fixed cost, value of the market) is ignored until Part B, which we will take up in the next chapter. By focusing on just this one unknown quantity for now, we can simplify the analysis as we introduce some of the most important fundamental ideas of probability theory.

2. Charting a probability distribution

We use probability distributions to describe people's beliefs about unknown quantities. When an unknown quantity has only finitely many possible values, we can describe it using a discrete probability distribution. (Continuous probability distributions, for unknown quantities with infinitely many possible values, will be discussed in the next chapter.) A discrete probability distribution can be presented in a table that lists the possible values of the unknown quantity and the probability of each possible value.

In the "Superior Semiconductors" case, the number of competitors who will enter the market is a quantity that is unknown to the decision-makers in this case, and they believe that this unknown quantity could be any number from 1 to 5. To use mathematical notation, let **K** denote this unknown number of competitors who will enter this market. (It is traditional among mathematicians to represent unknown quantities by boldface letters.) Then the decision-maker's beliefs about this unknown quantity **K** are described in the case by a discrete probability distribution such that

$$P(\mathbf{K}=1) = 0.10, \quad P(\mathbf{K}=2) = 0.25, \quad P(\mathbf{K}=3) = 0.30, \quad P(\mathbf{K}=4) = 0.25, \quad P(\mathbf{K}=5) = 0.10 .$$

Here for any number k , the mathematical expression $P(\mathbf{K}=k)$ denotes the probability that the unknown quantity **K** is equal to the value k . This probability distribution summarized by a chart in Figure 1.

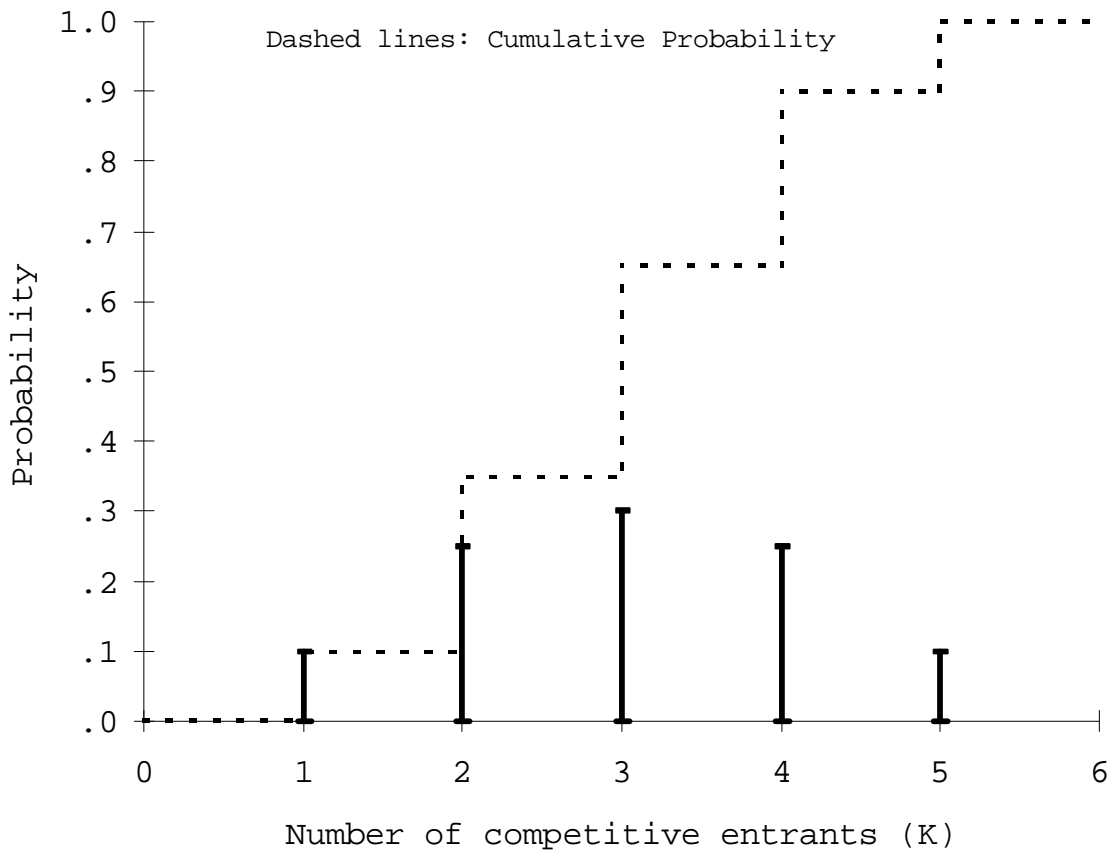


Figure 1. Discrete probability distribution for the number of entrants (K).

Figure 1 actually displays this probability distribution in two different ways. The five solid bars in Figure 1 show the probabilities of the five points on the horizontal axis that represent possible values of the unknown quantity \mathbf{K} . Such point-probability bars are the most common way of exhibiting a discrete probability distribution. But Figure 1 also contains a dashed line that shows the cumulative probabilities $P(\mathbf{K} < k)$ for each number k on the horizontal axis. For any number k between 1 and 2, the height of the dashed cumulative-probability curve in Figure 1 is 0.10, because $P(\mathbf{K} < k) = P(\mathbf{K} = 1) = 0.10$ when $1 < k < 2$. For any number k between 2 and 3, the height of the dashed cumulative-probability curve in Figure 1 is 0.35, because $P(\mathbf{K} < k) = P(\mathbf{K} = 1) + P(\mathbf{K} = 2) = 0.10 + 0.25 = 0.35$ when $2 < k < 3$. The unknown quantity \mathbf{K} is sure to be less than 6, and so the height of the cumulative probability-curve at 6 is $1 = P(\mathbf{K} < 6)$. The unknown quantity \mathbf{K} is sure to not be less than 0, and so the height of the cumulative-probability curve at 0 is

$$0 = P(\mathbf{K} < 0).$$

In Figure 1, notice that the vertical jumps in the dashed cumulative-probability curve occur exactly where the point-probability bars occur, and the height of each vertical jump is the same as the height of the corresponding point-probability bar. For example, the dashed cumulative-probability curve jumps from 0.10 to 0.35 above the value 2 on the horizontal axis of Figure 1, which corresponds to the fact that $P(\mathbf{K}=2) = 0.25 = 0.35 - 0.10$. So the cumulative-probability curve tells us everything about the probability distribution that we could learn from the point-probability bars. This observation is important, because we will find that cumulative-probability curves are generally more useful for describing probability distributions than point-probability bars (which cannot be applied to continuous probability distributions where there are infinitely many possible values).

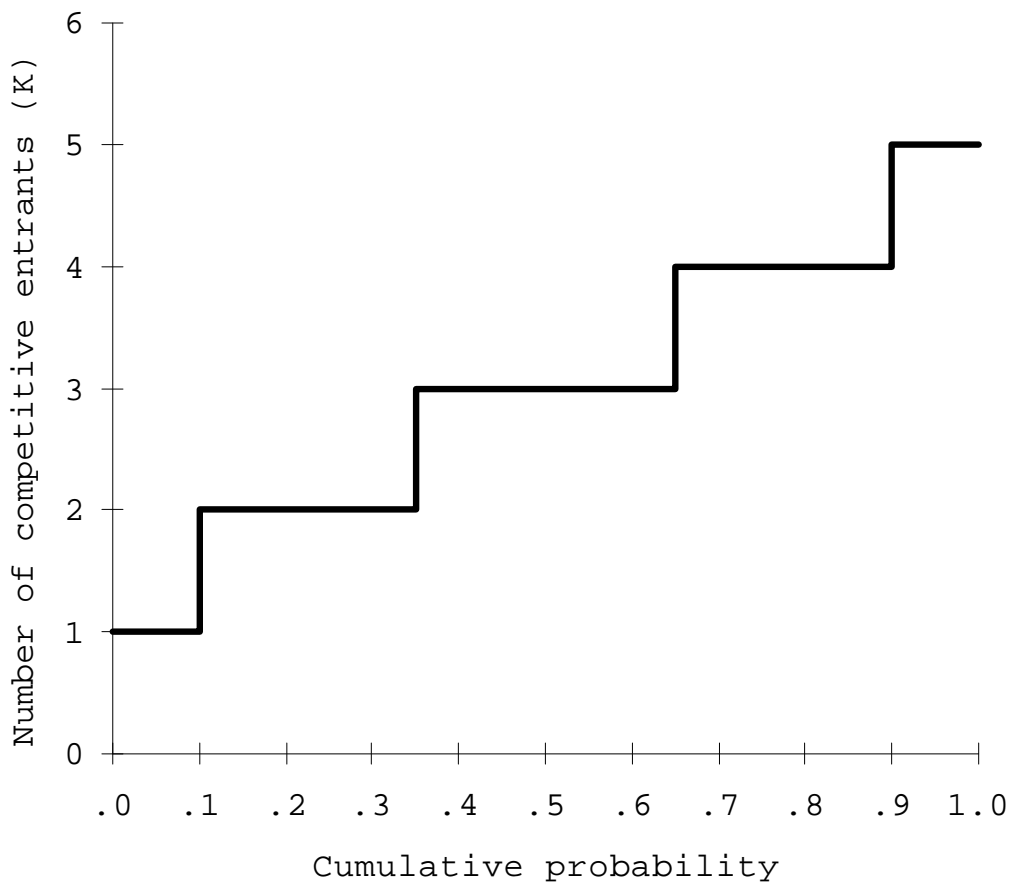


Figure 2. Inverse cumulative probability curve for the number of entrants (K).

Actually, as noted in Chapter 1, we will find it most useful to invert the cumulative probability distribution, turning the dashed line from Figure 1 on its side, with cumulative probabilities on the horizontal axis and possible values of \mathbf{K} on the vertical axis. Such an inverse cumulative-probability curve is shown in Figure 2. Once you learn how to read it, you can find the discrete probabilities of all possible values of \mathbf{K} from this inverse cumulative-probability curve. For example, the inverse cumulative-probability curve has height 2 over the interval of probabilities from 0.10 to 0.35, which tells us that the point-probability of the value 2 is $P(\mathbf{K}=2) = 0.35 - 0.10 = 0.25$. The height of the inverse cumulative-probability curve goes from 1 to 5 because these are the lowest and highest possible values of the unknown \mathbf{K} .

3. Simulation with discrete random variables

When we say that a random variable in a spreadsheet simulates (or represents) some unknown quantity in real life, we mean that any event for this simulated random variable is, from the perspective of our current information and beliefs, just as likely as the same event for the real unknown quantity. If the unknown number of competitive entrants has a probability 0.10 of being 1, for example, then the random variable in the spreadsheet should also have probability 0.10 of being 1 after the next recalculation of the spreadsheet. For any number k , the probability that the random variable will be less than k after the next recalculation should be the same as the probability that the real unknown quantity is less than k .

In our spreadsheets, all our random variables are constructed as functions of $\text{RAND}()$ values. So recall how the $\text{RAND}()$ function operates: Given any two numbers x and y such that $0 \leq x \leq y \leq 1$, the event that some particular $\text{RAND}()$ in a spreadsheet formula will take a value between x and y after the next recalculation is equal to the difference $y - x$. For example, the probability that a $\text{RAND}()$ will be between 0 and 0.10 is 0.10, and the probability that the $\text{RAND}()$ will be between 0.10 and 0.35 is $0.35 - 0.10 = 0.25$.

Let us now try to make a spreadsheet formula that depends on one $\text{RAND}()$ such that the value of our formula will simulate the unknown quantity \mathbf{K} from the "Superior Semiconductors" case. To be specific let us enter $=\text{RAND}()$ into cell A13, as shown in Figure 3. Into another cell,

we want to enter a formula that depends on the RAND() in cell A13 in such a way that its value is 1 with probability 0.10, 2 with probability 0.25, 3 with probability 0.30, 4 with probability 0.25, and 5 with probability 0.10, just like the unknown quantity **K**. One way to get these probabilities is to use some formula that returns a value that depends on the RAND() in cell A13 as follows:

- the value is 1 when A13 is less than 0.10 (an event having probability 0.10),
- the value is 2 when A13 is between 0.10 and $0.10+0.25=0.35$ (having probability 0.25),
- the value is 3 when A13 is between 0.35 and $0.35+0.30=0.65$ (having probability 0.30),
- the value is 4 when A13 is between 0.65 and $0.65+0.25=0.90$ (having probability 0.25),
- the value is 5 when A13 is greater than 0.90 (having probability 0.10).

Notice that we have gotten all the probabilities that we want with a collection of disjoint intervals that exhaust all the possible RAND() values because the probabilities of all possible values of **K** sum to 1. This is why all the possible values in any discrete probability distribution must have probabilities that sum to exactly 1!

There are several ways to write an Excel formula that will have these properties. One such formula is

$=1+IF(A13>=0.1,1,0)+IF(A13>=0.35,1,0)+IF(A13>=0.65,1,0)+IF(A13>=0.9,1,0)$

(Here ">=" means "greater than or equal to.") This long formula will do what we want, but it is not recommended. You should always try to avoid typing such a long formula into a spreadsheet, not only because it is tedious, but because your chances of making a typographical error somewhere in the formula are too great. So we need some shorter way of doing the same thing.

A better way to accomplish the same result is to use the Simtools function DISCRINV. This function takes three parameters. The first parameter, which is called "randprob" in the Insert-Function dialogue box, should simply be a RAND(). The second parameter, called "values" in the Insert-Function dialogue box, should be a range that lists the possible values of our discrete random variable. The third parameter, called "probabilities" in the Insert-Function dialogue box, should be another range which has the same size as the values range and which lists the corresponding probabilities of these values. Then the formula

$DISCRINV(RAND(), \text{values}, \text{probabilities})$

returns a random variable that has values and discrete probabilities as listed in these ranges.

	A	B	C	D	E	F	G	H
1	A DISCRETE RANDOM VARIABLE FROM THE "SUPERIOR SEMICONDUCTOR" CASE							
2		K = (unknown number of competitors entering market).						
3		Little k = (possible value of big K).						
4	P(K<k)	k	P(K=k)					
5	0.00	1	0.10					
6	0.10	2	0.25					
7	0.35	3	0.30					
8	0.65	4	0.25					
9	0.90	5	0.10					
10			1					
11								
12	(rand)	Simulated value						
13	0.315424	2						
14		2						
15		2						
16								
17	FORMULAS FROM RANGE A1:C15							
18	A6.	=A5+C5						
19	A6 copied to A6:A9							
20	C10.	=SUM(C5:C9)						
21	A13.	=RAND()						
22	B13.	=1+IF(A13>=0.1,1,0)+IF(A13>=0.35,1,0)+IF(A13>=0.65,1,0)+IF(A13>=0.9,1,0)						
23	B14.	=DISCRINV(A13,\$B\$5:\$B\$9,\$C\$5:\$C\$9)						
24	B15.	=VLOOKUP(A13,A5:B9,2)						

Figure 3. Simulation with a discrete probability distribution (three equivalent formulas).

In Figure 3, for example, the possible values of the unknown quantity **K** {1;2;3;4;5} are listed in the range B5:B9, while the corresponding probabilities {0.10;0.25;0.30;0.25;0.10} are listed in the range C5:C9. So to simulate the unknown quantity **K** in our spreadsheet, we can use the formula

$$=DISCRINV(A13, B5:B9, C5:C9)$$

The results of this Simtools function are exactly the same as the long formula shown above: returns the value 1 when $A13 < 0.10$, it returns the value 2 when $0.10 \leq A13 < 0.25$, and so on. In Figure 3, for example, that the value of cell A13 (0.315) happens to be between 0.10 and 0.35, and so the value of this DISCRINV formula in cell B14 is 2.

At this point you may wonder why this function should be called DISCRINV. This function has been designed for simulating a discrete random variable, and so the letters "DISC" obviously come from the word "discrete," But what does the "INV" signify? To see the answer,

notice that this function has been designed to return the value 1, 2, 3, 4, or 5, depending on the value of the RAND() that is its first parameter, and the points where the function's value changes are 0.10, 0.10+0.25 = 0.35, 0.35+0.30 = 0.65, and 0.65+0.25 = 0.90. You have seen a function like this before: it is the inverse cumulative function shown in Figure 2! Indeed, the "INV" in our function's name is short for "inverse cumulative."

In general, any random variable can be simulated by processing a RAND() through the inverse cumulative-probability function. To say that a function $G(\bullet)$ is the inverse cumulative-probability function of an unknown quantity \mathbf{X} is to say that $P(\mathbf{X} < G(p)) = p$ for every number p between 0 and 1. But then $G(\text{RAND}())$ is less than $G(p)$ when the $\text{RAND}()$ is less than p , which happens with the same probability p . So $G(\text{RAND}())$ has the same probability distribution as \mathbf{X} .

(If you do not have Simtools.xla, then the values returned by DISCRINV can also be returned by using Excel's VLOOKUP function as shown in Figure 3. The trick involves computing the strict cumulative probabilities for the various possible values in a column to the left of the possible values, as shown in the range A5:A9 of Figure 3. Then VLOOKUP can be used with a RAND() as its first parameter, the range from cumulative probabilities to the corresponding values as its second parameter, and the column-number of the values in this range as its third parameter.)

The bottom-line quantity of interest in the "Superior Semiconductors" is profit, which is assumed to depend on the number of competitors \mathbf{K} by the formula

$$\text{Profit} = 100 / (1 + \mathbf{K}) - 26$$

So if \mathbf{K} equals 1 then profit is $100 / (1 + 1) - 26 = 24$ (in \$millions), but if \mathbf{K} equals 5 then profit is $100 / (1 + 5) - 26 = -9.33$. So profit is also an unknown quantity, with a discrete probability distribution as shown in the following table:

Competitors	Profit	Probability
1	24	0.10
2	7.33	0.25
3	-1	0.30
4	-6	0.25
5	-9.33	0.10

Figure 4 shows two attempts to make a model in which the number of competitors and the

resulting profit for Superior Semiconductors are both simulated.

	A	B	C	D	E	F	G	
1	"SUPERIOR SEMICONDUCTOR" CASE				26	FixedCost		
2					100	MarketValue		
3	Let K = (unknown number of competitors entering market)							
4		k	P(K=k)		Profit	(in \$millions)		
5		1	0.10		24.00			
6		2	0.25		7.33			
7		3	0.30		-1.00			
8		4	0.25		-6.00			
9		5	0.10		-9.33			
10			1	sum				
11								
12		Model 1 (correct)						
13		#Competitors entering			Profit			
14		2			7.33			
15								
16		Model 2 (WRONG!!!)						
17		#Competitors entering			Profit			
18		4			24			
19								
20	FORMULAS FROM RANGE A1:F18							
21	E5.	=E\$2/(1+B5)-E\$1						
22	E5	copied to E5:E9						
23	C10.	=SUM(C5:C9)						
24	B14.	=DISCRINV(RAND(),B5:B9,\$C\$5:\$C\$9)						
25	E14.	=E\$2/(1+B14)-E\$1						
26	B18.	=DISCRINV(RAND(),B5:B9,\$C\$5:\$C\$9)						
27	E18.	=DISCRINV(RAND(),E5:E9,\$C\$5:\$C\$9)						

Figure 4. Making a simple simulation model of competitors and profit.

The range B16:E18 in Figure 4 contains an attempt, called "Model 2," which illustrates one of the most common errors that students make in simulation modeling. In Model 2, the number of competitors and the profit are simulated in cells B18 and E18 respectively, with two DISCRINV formulas that depend on separate RANDs. The result is that profit is independent of the number of competitors in Model 2, which is wrong! For example, Figure 4 shows a realization of these random variables such that the simulated number of competitors in B18 is 4 while the simulated profit in E18 is 24 (\$millions); but if there were really 4 competitive entrants then Superior Semiconductor's profit would be $100/(1+4)-26 = -6$.

The range B12:E14 in Figure 4 contains "Model 1," which is done correctly. In this model, the number of competitors is simulated by a DISCRINV formula in cell B14, while the profit is simulated in cell E14 as a function of the simulated number of competitors by the formula $=E2/(1+B14)-E1$ (where E2 contains the market value 100 and E1 contains the fixed cost 26). Thus Model 1 always displays the correct relationship between the number of competitors and the profit.

In general, a simulation model is a good representation of a real situation if our uncertainty about the next recalculated values of the random variables in the model is the same as our uncertainty about the corresponding unknown quantities in the real situation. Fancy formulas in a spreadsheet obviously cannot be asked to magically return the actual values of real-world quantities that we are unable to observe or measure by other means. Instead, what we must ask of our simulation models is that their formulas should express our beliefs about the real unknown quantities, in the sense that our beliefs about the next recalculated values of these formulas are the same as our beliefs about the real unknown quantities.

4. Expected value and standard deviation

We have seen how a discrete probability distribution can be used to describe our beliefs about some unknown quantity that has finitely many possible values, how such probability distributions can be exhibited graphically, and how to make random variables in a spreadsheet for simulating such probability distributions. But when there are many possible values, a probability distribution may be quite complicated. In such cases, we may want to describe the overall pattern of a probability distribution by a few summary numbers which people could interpret more easily than some complicated a chart or some simulated random variable that jumps around whenever [F9] is pressed.

There are many formulas that people have used to generate summary measures of probability distributions (expected value, median, mode, standard deviation, mean absolute error, etc.). Each of these formulas has some drawbacks and limitations, because it is impossible to perfectly summarize everything we want to know about every probability distribution by just a couple of simple numbers. But two summary measures have been found particularly useful and

will be emphasized in this course: the expected value and the standard deviation.

The mean or expected value of an unknown quantity \mathbf{X} may be denoted by $E(\mathbf{X})$ or $\mu_{\mathbf{X}}$, and it is defined by the formula

$$E(\mathbf{K}) = \mu_{\mathbf{X}} = \sum_x P(\mathbf{X}=x)*x$$

where the summation is over all numbers x that are possible values of the unknown quantity \mathbf{X} . For example, in the "Superior Semiconductor" case, the expected value of the unknown number of competitors \mathbf{K} is

$$E(\mathbf{K}) = 0.10*1 + 0.25*2 + 0.30*3 + 0.25*4 + 0.10*5 = 3.$$

Similarly, if we let $\mathbf{Y} = 100/(1+\mathbf{K})-26$ denote the profit in this case, then the expected value of profit is

$$E(\mathbf{Y}) = 0.10*24 + 0.25*7.33 + 0.30*(-1) + 0.25*(-6) + 0.10*(-9.33) = 1.5.$$

These calculations are illustrated in cells B13 and F13 of Figure 5, where the important Excel function SUMPRODUCT is used. When "range1" and "range2" denote two ranges that have the same numbers of rows and columns in a spreadsheet, the Excel formula SUMPRODUCT(range1,range2) multiplies the values of each pair of corresponding cells in this ranges (starting with the top-left cell of range1 multiplied by the top-left cell in range 2) and then adds up all of these products. Thus, when the possible values of \mathbf{K} are listed in the range B5:B9 and the corresponding probabilities are listed in C5:C9, the expected number of competitors $E(\mathbf{K})$ can be returned by the formula

$$=SUMPRODUCT(B5:B9,C5:C9)$$

in cell B13. Similarly, when the corresponding profit levels in are computed in the range F5:F9, the expected profit $E(\mathbf{Y})$ can be returned by the formula

$$=SUMPRODUCT(F5:F9,C5:C9)$$

in cell F13 of Figure 5.

	A	B	C	D	E	F	G	H
1	DISCRETE RANDOM VARIABLES FROM THE "SUPERIOR SEMICONDUCTOR" CASE							
2	K = (unknown number of competitors entering market).							
3	Little k = (possible value of big K).							
4		k	P(K=k)	(k-E(K))^2		Y=Profit		
5		1	0.1	4		24		
6		2	0.25	1		7.333333		
7		3	0.3	0		-1		
8		4	0.25	1		-6		
9		5	0.1	4		-9.333333		
10			1					
11								
12		Mean or E(K)	Stdev(K)			E(Y)	Stdev(Y)	
13		3	1.14018	1.140175		1.5	9.31695	
14								
15	FORMULAS FROM RANGE A1:G13							
16	D5.	=(B5-\$B\$13)^2						
17	D5 copied to D5:D9							
18	F5.	=100/(1+B5)-26						
19	F5 copied to F5:F9							
20	C10.	=SUM(C5:C9)						
21	B13.	=SUMPRODUCT(B5:B9,\$C\$5:\$C\$9)						
22	C13.	=STDEVPR(B5:B9,\$C\$5:\$C\$9)						
23	D13.	=SUMPRODUCT(D5:D9,C5:C9)^0.5						
24	F13.	=SUMPRODUCT(F5:F9,\$C\$5:\$C\$9)						
25	G13.	=STDEVPR(F5:F9,\$C\$5:\$C\$9)						

Figure 5. Expected values and standard deviations of discrete random variables.

The expected value is intended to be a measure of the center of a probability distribution. There will generally be some possible values that are higher than the expected value and other possible values that are lower.

Notice, however, that the expected value of an unknown quantity is not necessarily itself a possible value of the unknown quantity. In this example, the expected value of **K** ($E(\mathbf{K})=3$) happens to be a possible value of **K**. But the expected value of **Y** ($E(\mathbf{Y})=1.5$) is not among the possible values of the unknown profit **Y**. So the term "expected" is being used here in a technical sense that may be different from common English usage of the word. In each case, however, the expected value could be reasonably described (in some intuitive sense) as "near the center" of the possible values of the unknown quantity.

Notice also that the expected profit in this example is different from the profit that occurs

at the expected number of competitors, even though profit here is a function that depends on the number of competitors. When the number of competitors \mathbf{K} is 3, which is $E(\mathbf{K})$, the corresponding profit is $\mathbf{Y} = 100/(1+3)-26 = -1$, but $E(\mathbf{Y}) = 1.5$. More generally, if \mathbf{X} is an unknown quantity and $f(\bullet)$ is any function, the function evaluated f at $E(\mathbf{X})$ may be different from the expected value of the unknown quantity $f(\mathbf{X})$, that is:

$$f(E(\mathbf{X})) \text{ may be different from } E(f(\mathbf{X})).$$

The erroneous assumption that $f(E(\mathbf{X}))$ ought to be the same as $E(f(\mathbf{X}))$ has been called the "flaw of averages" (in echo of the more respectable "law of averages" that we will discuss soon). This error seems to arise in people's minds because, after computing the expected value of an unknown quantity, there is a temptation to simplify the world by assuming that the unknown quantity will be equal to its expected value. For example, in the "Superior Semiconductor" case, we might be tempted to assume that the number of competitors will be 3 for sure, in which case the profit would be $100/(1+3)-26 = -1$ (a loss of \$1 million). But when the uncertainty is not assumed away we actually get a positive expected profit of 1.5 \$million. This positive expected value of profits should seem intuitively reasonable when you notice that the positive profits (24 and 7.33) generated by numbers of competitors below 3 are significantly larger in absolute value than the equally-likely negative profits (-9.33 or -6) generated by numbers of competitors above 3. This result is in turn caused by a kind of nonlinearity in our profit function (where a decrease of \mathbf{K} below 3 would increase profit by more than the a similar increase of \mathbf{K} above 3 would decrease profit).

Now that we have the expected value as a summary measure of the center of a probability distribution, we should also want some summary measure of the spread of a probability distribution, to say something about what kinds of deviations from this expected value are likely to occur. The most useful summary measure of spread is the standard deviation.

The standard deviation of a random variable \mathbf{X} may be denoted by $\text{Stdev}(\mathbf{X})$ or $\sigma_{\mathbf{X}}$, and it is defined by the formula

$$\text{Stdev}(\mathbf{X}) = \sigma_{\mathbf{X}} = (E((\mathbf{X}-\mu_{\mathbf{X}})^2))^{0.5} = (\sum_{\mathbf{x}} P(\mathbf{X}=\mathbf{x})*(\mathbf{x}-\mu_{\mathbf{X}})^2)^{0.5}$$

(Here $\mu_{\mathbf{X}} = E(\mathbf{X})$, and the summation ($\sum_{\mathbf{x}}$) is over all numbers \mathbf{x} that are possible values of the unknown quantity \mathbf{X} . The symbol " \wedge " is used in here and in Excel to indicate exponentiation, so

$3^2 = 3^2 = 9$ for example.) In words, the standard deviation of \mathbf{X} may be defined as the square root of the expected squared deviation of \mathbf{X} from its mean. If we did drop square root (" $\wedge 0.5$ ") from this definition, then we get the definition of the variance of \mathbf{X} , which is the expected value of the squared deviation of \mathbf{X} from its mean

$$\text{Var}(\mathbf{X}) = E((\mathbf{X} - \mu_{\mathbf{X}})^2) = \sum_x P(\mathbf{X}=x) * (x - \mu_{\mathbf{X}})^2 = (\text{Stdev}(\mathbf{X}))^2$$

In the "Superior Semiconductor" example, the unknown number of competitors \mathbf{K} has standard deviation

$$(0.10*(1-3)^2 + 0.25*(2-3)^2 + 0.30*(3-3)^2 + 0.25*(4-3)^2 + 0.10*(5-3)^2)^{0.5} = 1.14$$

To see how these calculations may be done in a spreadsheet, look at cells D5:D9 and D13 in Figure 5. Recall that the possible values of \mathbf{K} are listed in cells B5:B9, the corresponding probabilities are listed in C5:C9, and the expected value or mean of \mathbf{K} is listed in cell B13 of Figure 5. So entering the formula

$$=(B5 - \$B\$13)^2$$

into cell D5, and copying D5 to D5:D9, we get the possible squared deviations of \mathbf{K} from its mean listed in cells D5:D9. Then the formula `SUMPRODUCT(D5:D9,C5:C9)` returns the expected squared deviation of \mathbf{K} from its mean. This expected squared deviation of a random variable from its mean is called the variance of the random variable. The standard deviation is the square root of the variance, and it is returned in cell D13 by the formula

$$=SUMPRODUCT(D5:D9,C5:C9)^{0.5}$$

To simplify these calculations, we may also use a function `STDEVPR` which is added by Simtools. This function is designed to compute standard deviations from discrete probability distributions. (Be careful not to be confuse Simtools's `STDEVPR` function with Excel's functions `STDEV` and `STDEVP`, which are used to compute sample standard deviations from sample data. The "PR" in `STDEVPR` is short for "PRobability distribution.") `STDEVPR` takes two parameters, which should be ranges that have the same numbers of rows and columns. The first parameter, called "values" in the Insert-Function dialogue box, should be a range that lists the possible values of some discrete random variable. The second parameter, called "probabilities" in the Insert-Function dialogue box, should be a range which lists the corresponding probabilities of these values. Then the formula `STDEVPR(values, probabilities)` returns the standard deviation of

the discrete random variable. For example the formula

$$=STDEVPR(B5:B9,C5:C9)$$

in cell C13 of Figure 5 returns the same standard deviation (1.14) that we computed in cell D13.

If you have never seen a standard deviation before, then I have to tell you that learning to interpret standard deviations takes time. There is no simple rule about how to interpret them. The most that I can say now is that the unknown quantity generally has some substantial probability of being more than one standard deviation above or below its expected value, but it is generally very unlikely to be more than three standard deviations above or below its expected value. Thus, for example, if you only told me that $E(\mathbf{K}) = 3.0$ and $Stdev(\mathbf{K}) = 1.14$ for some unknown quantity \mathbf{K} , then I would infer that the probability of \mathbf{K} being above $3.0 + 1.14 = 4.14$ or below $3.0 - 1.14 = 1.86$ was nonnegligible (this probability is actually 0.20 in our example), but the probability of \mathbf{K} being above $3.0 + 3 * 1.14 = 6.42$ or below $3.0 - 3 * 1.14 = -0.42$ was very small (this probability is actually 0 in our example). If these vague inferences were not enough for me, then I would ask you to show me the (inverse) cumulative chart for this unknown quantity.

There are two formulas that you should know about how multiplication by a nonrandom constant c and addition of a nonrandom constant d would affect expected values and standard deviations. If \mathbf{X} is a random variable but c and d are numbers which are not random, then

$$E(c * \mathbf{X} + d) = c * E(\mathbf{X}) + d,$$

$$Stdev(c * \mathbf{X} + d) = |c| * Stdev(\mathbf{X}).$$

(In the second formula, $|c|$ denotes the absolute value of c , that is, $|c| = c$ if $c \geq 0$, while $|c| = -c$ if $c < 0$.) For example, let the random variable \mathbf{R} denote the Superior Semiconductor's total revenue in dollars (not \$millions) from the new product, ignoring the fixed costs. This revenue number \mathbf{R} in dollars and the profit \mathbf{Y} in millions of dollars are obviously related by the formula $\mathbf{R} = 1,000,000 * \mathbf{Y} + 26,000,000$. So knowing that $E(\mathbf{Y}) = 1.5$ and $Stdev(\mathbf{Y}) = 9.317$, we can compute

$$E(\mathbf{R}) = 1,000,000 * E(\mathbf{Y}) + 26,000,000 = 27,500,000$$

$$Stdev(\mathbf{R}) = 1,000,000 * Stdev(\mathbf{Y}) = 9,317,000.$$

5. Estimates from sample data

In complex decision problems, we will want to estimate the expected values, standard deviations, and cumulative distributions of unknown quantities that we can study only by simulation. In this section we show how such estimates can be generated. It will be helpful to begin studying these techniques in the context of the simple "Superior Semiconductors (A)" example, where we know how to explicitly compute the numbers that we are trying to estimate from simulation, because this simplicity will enable you to get a hands-on feel for the accuracy of these simulation methods.

Figure 6 shows a table of simulations of the unknown number of competitors **K** from the "Superior Semiconductors (A)" case. The possible values of **K** are listed in cells A3:A7, and their corresponding probabilities are listed in cells B3:B7. So of course we can compute the expected value and standard deviation of **K** directly from this probability distribution, as shown in cells D7 and E7 of Figure 6. But let us pretend that we did not know how to do these computations and instead try to work with simulation data.

To make the simulation data in this spreadsheet, I entered into cell B14 the formula

`=DISCRINV(RAND(),A3:A7,B3:B7)`

to make a random variable that simulates the unknown quantity **K**. Next, I selected the range A14:A415 and entered the command sequence Tools>SimTools>SimulationTable. As a result, the range B15:B415 was filled with the values of 401 independent recalculations of the random variable in cell B14. Also, the label "SimTable" was entered into cell A14, and the range A15:A415 was filled with a percentile index consisting of 401 equally-spaced numbers from 0 to 1.

To remind us how many data points we have in our simulation data (401), I entered the formula `=COUNT(B15:B415)` into cell B11. Notice that the data range, as defined here, does not include the original random variable in cell B14. If we had included it, the statistics that we compute would change slightly every time the spreadsheet was recalculated. In our mathematical formulas, the number of independent data points that we are using to compute our statistics will be commonly denoted by the letter "n".

	A	B	C	D	E	F	G	H	I
1	Probability distribution of K (#competitors entering market)								
2	k	P(K=k)							
3	1	0.1							
4	2	0.25							
5	3	0.3							
6	4	0.25		E(K)	Stdev(K)				
7	5	0.1		3	1.14018				
8									
9		3.07481	E(K) estimated from Simtable						
10		1.12445	Stdev(K) estimated from Simtable						
11		401	Sample size n						
12									
13		Sim'd K:	1.12445	(SumSqDevs/(n-1))^0.5					
14	SimTable	3	Squared deviations from sample mean						
15	0	4	0.85597						
16	0.0025	4	0.85597		FORMULAS FROM RANGE A1:F15				
17	0.005	3	0.00560		D7. =SUMPRODUCT(A3:A7,B3:B7)				
18	0.0075	2	1.15522		E7. =STDEVPR(A3:A7,B3:B7)				
19	0.01	1	4.30485		B9. =AVERAGE(B15:B415)				
20	0.0125	4	0.85597		B10. =STDEV(B15:B415)				
21	0.015	3	0.00560		B11. =COUNT(B15:B415)				
22	0.0175	5	3.70635		D13. =(SUM(D15:D415)/(B11-1))^0.5				
23	0.02	3	0.00560		B14. =DISCRINV(RAND(),A3:A7,B3:B7)				
24	0.0225	2	1.15522		D15. =(B15-\$B\$9)^2				
25	0.025	2	1.15522		D15 copied to D15:D415				

Figure 6. Estimating an expected value and standard deviation from simulation data.

The best estimate of $E(\mathbf{K})$ that can be computed from a sample of independent simulated values of \mathbf{K} is the average or sample mean of this simulation data. So to estimate $E(\mathbf{K})$ in Figure 6, the formula

$$=AVERAGE(B15:B415)$$

has been entered into cell B9. Excel's AVERAGE function, of course, just sums the n numerical values in our data range and divides this sum by n . You can see that the value returned in this case (3.07481) is reasonably close to the actual expected value (3).

The law of large numbers is a mathematical theorem which asserts that, when we have a very large number of values drawn independently from a fixed probability distribution, the average of these values is very likely to be quite close to the expected value of the probability distribution. (I am giving you just an informal description of this theorem here. Its formal mathematical statement gives precise meanings to my phrases "very large", "very likely", and "quite close".) To

see why this law is true, consider any discrete probability distribution like the one for \mathbf{K} here. If we let m_k denote the number of times that the value k occurs in our data range and let n denote the size of the whole data range, then the average of the data range is

$$(\sum_k k*m_k)/n = \sum_k k*(m_k/n)$$

where the summation is over all numbers k that are possible values of the random variable. But when we generate hundreds of independent samples from the probability distribution of \mathbf{K} , we should anticipate that the relative frequency m_k/n of each possible value k should be close to its probability. That is, m_k/n should be quite close to $P(\mathbf{K}=k)$. Thus, the average of the data range should be close to the expected value

$$E(\mathbf{K}) = \sum_k k*P(\mathbf{K}=k).$$

For example, in the spreadsheet shown in Figure 6, the value 1 actually occurs 34 times among the 401 data values in the range B15:B415 (most of which are not shown in the figure). So $m_1/n = 34/401 = 0.085$, which is not far from $0.10 = P(\mathbf{K}=1)$. The other frequencies in this data set happen to be such that

$$\begin{aligned} (\sum_k k*m_k)/n &= (1*34 + 2*98 + 3*112 + 4*118 + 5*39)/401 \\ &= \sum_k k*m_k/n = 1*34/401 + 2*98/401 + 3*112/401 + 4*118/401 + 5*39/401 \\ &= 1*0.085 + 2*0.244 + 3*0.279 + 4*0.294 + 5*0.097 = 3.07481 \end{aligned}$$

which is the value shown in cell B9 of Figure 6. Matching the actual relative frequencies with the theoretical probabilities, you can see why this average is close to the true expected value

$$E(\mathbf{K}) = \sum_k k*P(\mathbf{K}=k) = 1*0.10 + 2*0.25 + 3*0.30 + 4*0.25 + 5*0.10 = 3.$$

To estimate the standard deviation from sample data, we use the Excel function STDEV. For example, cell B10 of Figure 6 contains the formula

$$=STDEV(B15:B415)$$

which returns the value 1.12445. You can see that this estimate is not far from the true standard deviation shown in cell E7 (1.14018).

Cells D13:D415 in Figure 6 have been set up to show you how Excel's STDEV function really works. Recall that the standard deviation of a random variable is the square root of its variance, and its variance is the expected squared deviation of this random variable from its expected value. So to estimate the standard deviation a random variable from sample data,

Excel's STDEV function must begin by estimating the expected value, which it does by computing the sample average, just as we have done in cell B9. Next, the STDEV function computes the squared deviation of each value in the data range from this sample average. These squared deviations have been computed in Figure 6 by entering the formula

$$=(B15-\$B\$9)^2$$

into cell D15, and then copying D15 to cells D15:D415. Next, you might think that we should estimate the variance (which is the expected squared deviation) by computing the average of these squared deviations. But statisticians have recommended instead that at this step we should instead compute a modified average in which the sum of the squared deviations is divided by $n-1$ instead of n . (Computing deviations from the sample average instead of the true expected value tends to slightly reduce the average of these squared deviations, which could cause a downward bias in our estimates of the variance. To correct for this downward bias, statisticians have recommended dividing by $n-1$ in the variance estimator.) In Figure 6, this modified average of the squared deviations would be returned by the formula

$$\text{SUM}(D15:D415)/(B11-1)$$

(Recall that the sample size n is in cell B11.) Finally, to reverse the squaring of the deviations, STDEV takes the square root of this sample variance estimate. Taking the square root is the same as raising to the 0.5 power, so the formula

$$=(\text{SUM}(D15:D415)/(B11-1))^{0.5}$$

has been entered into cell D13 in Figure 6. You can verify that the result of this formula is exactly the value returned by the STDEV function (1.12445). This estimated standard deviation computed by STDEV from sample data is called the sample standard deviation.

The law of large numbers can be extended to these estimated sample standard deviations. That is, when we have a very large number of values drawn independently from a fixed probability distribution, the sample standard deviations of these values is very likely to be quite close to actual standard deviation of the probability distribution.

(Microsoft has provided STDEV as a built-in Excel function, because practical statistical work often requires that standard deviations be estimated from data. But the STDEVPR function is provided only by the Simtools.xla add-in because, in practice, standard deviations are only

rarely computed from the probability distributions that are define them.)

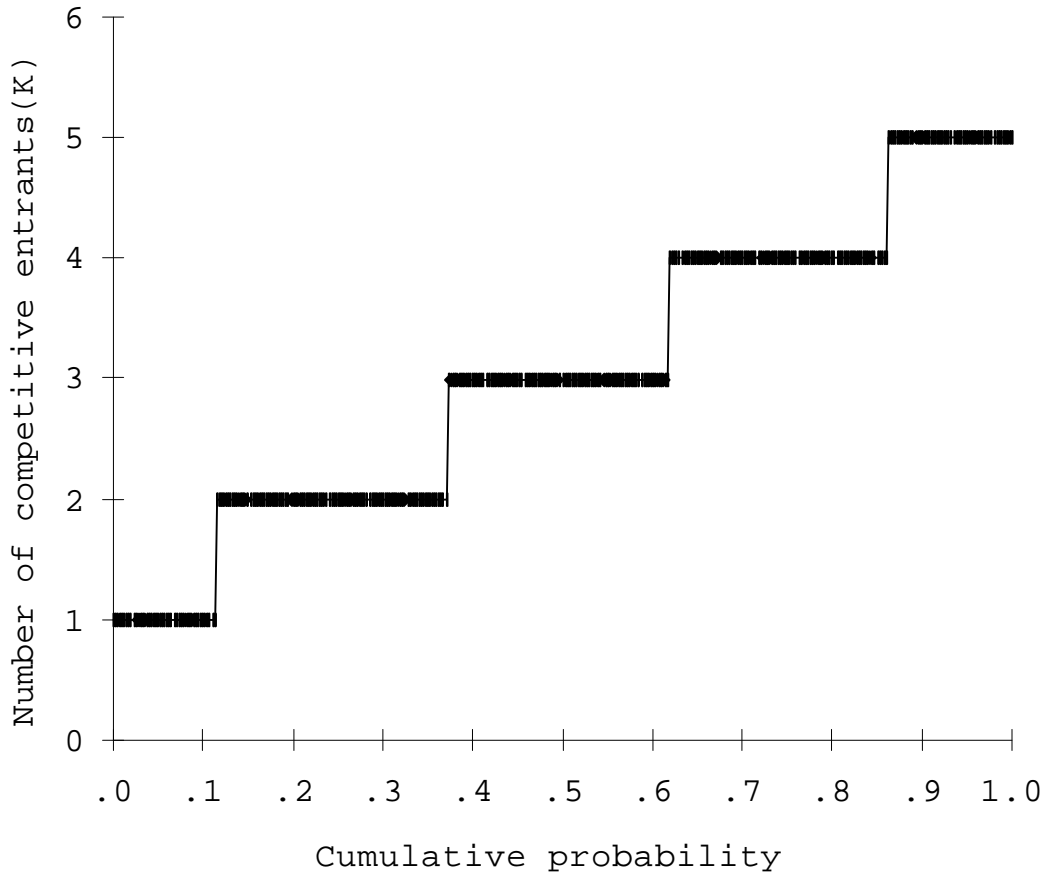


Figure 7. Estimate of an inverse cumulative distribution from simulation data.

To estimate the inverse cumulative distribution of a random variable from data in a simulation table, we only need to sort the simulated values in our data range (ascending) and plot these sorted on the vertical (Y) axis of in an XY-scatter chart, with the corresponding percentile-index values (from the left column of the simulation table) on the horizontal (X) axis of the chart. Figure 7 shows such a chart, made from the simulation table in Figure 6. To make Figure 7, I first selected the simulated data range B15:B415, and then I entered the command sequence Data>Sort>Ascending. (When Excel's "sort warning" dialogue box interrupted the task, I chose

the "continue" option.) Next, I selected the range A15:B415, which includes the percentile index numbers that go from 0 to 1 as well as the now-sorted data range, and then I entered the command sequence Insert>Chart>XY-scatter (selecting the option to show lines and point markers), to create the chart as shown in Figure 7. Notice that Figure 7 is indeed a good approximation to the actual inverse cumulative-probability chart shown previously as Figure 2.

6. Accuracy of sample estimates

When we estimate an expected value by computing the average of sample data, we need to know something about how accurate this estimate is likely to be. Of course the average of several random variables is itself a random variable that has its own probability distribution. Figure 8 shows a spreadsheet designed to help you learn about how sample averages behave as random variables. Like other figures in this chapter, Figure 8 begins with the probability distribution for the unknown quantity **K** from "Superior Semiconductors (A)", with the possible values listed in cells A3:A7 and their corresponding probabilities listed in cells B3:B7. Then the formula =DISCRINV(RAND(), \$A\$3:\$A\$7, \$B\$3,\$B\$7) has been entered into every cell in the range A15:C24, giving us 30 independent random variables drawn from the same probability distribution. The average or sample mean of these 30 random variables is entered into cell E15 by the formula =AVERAGE(A15:C24).

If you should make a spreadsheet like Figure 8, and recalculate it many times, watching the sample mean in cell E15. If you watch any individual cell in the sample range A15:C24, you will see it jump around to all the integer values from 1 to 5. But when you watch cell E15, the average of all 30 cells in this sample, you will see it vary much less widely around 3, almost never going below 2.2 or above 3.8. But a remarkable fact called the "central limit theorem" tells us much more about the way this average varies.

	A	B	C	D	E	F	G	H
1	Probability distribution of K (#competitors entering market)							
2	k	P(K=k)						
3	1	0.1						
4	2	0.25						
5	3	0.3						
6	4	0.25		E(K)		Stdev(K)		
7	5	0.1		3		1.140175		
8		1						
9				Sample size n		Stdev of sample mean		
10				30		0.208167		
11								
12	Normal random variable with same E & Stdev as sample mean: 2.886765							
13								
14	Simulated sample, size 30							
15	4	2	3		3.133333	Sample mean or average		
16	2	1	5		1.166585	Sample stdev		
17	3	4	3		0.212988	Est'd stdev of sample mean		
18	5	2	2					
19	2	2	3	95% confidence interval for E(K)				
20	4	5	1		2.715876	3.55079		
21	3	3	4	E(K) actually in the interval?				
22	4	3	2		TRUE			
23	3	2	4					
24	4	5	4					
25								
26	FORMULAS FROM RANGE A1:H24							
27	D7.	=SUMPRODUCT(A3:A7,B3:B7)			E15.	=AVERAGE(A15:C24)		
28	F7.	=STDEVPR(A3:A7,B3:B7)			E16.	=STDEV(A15:C24)		
29	B8.	=SUM(B3:B7)			E17.	=E16/(D10^0.5)		
30	D10.	=COUNT(A15:C24)			E20.	=E15-1.96*E17		
31	F10.	=F7/(D10^0.5)			F20.	=E15+1.96*E17		
32	H12.	=NORMINV(RAND(),D7,F10)			E22.	=AND(E20<D7,D7<F20)		
33	A15.	=DISCRINV(RAND(),\$A\$3:\$A\$7,\$B\$3:\$B\$7)						
34	A15 copied to A15:C24							

Figure 8. A spreadsheet for studying the properties of a sample mean.

Before stating the central limit theorem, I must introduce the idea of a Normal probability distribution. In probability theory, the phrase "Normal probability distribution" is used in a technical sense, which I will emphasize by capitalizing it, referring to a particular collection of mathematical probability distributions that have some important properties. For any two numbers μ and σ such that $\sigma > 0$, there is precisely-defined Normal probability distribution that has mean (or expected value) μ and standard deviation σ . We will discuss such Normal distributions at length in chapter 3, but for now it is enough to introduce them by citing a few basic facts about them.

In an Excel spreadsheet, you can make a random variable that has a Normal probability distribution with mean μ and standard deviation σ by the formula

$$=\text{NORMINV}(\text{RAND}(),\mu,\sigma)$$

I have tried to suggest here that, once you know how to make a spreadsheet cell that simulates a given probability distribution, you can learn anything that anybody might want to know about this distribution by simulating it many times in a spreadsheet. So if you want to know what a "Normal distribution with mean 100 and standard deviation 20" is like, you should simply copy the formula `=NORMINV(RAND(),100,20)` into a large range of cells and watch how the values of these cells jump around whenever you press the recalc key [F9].

Let me give you now a few other useful formulas about Normal distributions. If a random variable \mathbf{X} has a Normal probability distribution with mean μ and standard deviation σ (where μ and σ are given numbers such that $\sigma > 0$), then

$$P(\mathbf{X} < \mu) = 0.5 = P(\mathbf{X} > \mu),$$

$$P(\mu - \sigma < \mathbf{X} < \mu + \sigma) = 0.683$$

$$P(\mu - 1.96 * \sigma < \mathbf{X} < \mu + 1.96 * \sigma) = 0.95$$

$$P(\mu - 3 * \sigma < \mathbf{X} < \mu + 3 * \sigma) = 0.997$$

That is, a Normal random variable is equally likely to be above or below its mean, it has probability 0.683 of being less than one standard deviation away from its mean, it has probability 0.997 (almost sure) of being less than 3 standard deviations of its mean. And for constructing 95% confidence intervals, we will use the fact that a Normal random variable has probability 0.95 of being within 1.96 standard deviations from its mean.

Now we are ready for the remarkable central limit theorem, which tells us Normal distributions can be used to predict the behavior of sample averages.

Consider the average of n random variables that are drawn independently from a probability distribution with expected value μ and standard deviation σ . This average, as a random variable, has expected value μ , has standard deviation $\sigma / (n^{0.5})$, and has a probability distribution that is approximately Normal.

For example, cell E15 of Figure 8 contains the average of $n=30$ independent random variables that are drawn from a probability distribution which has expected value $\mu=3$ and standard deviation σ

= 1.14. So the central limit theorem tells us that this sample mean should behave like a random variable that has a Normal distribution where $\mu=3$ is the mean and $\sigma/(n^{0.5}) = 1.14/(30^{0.5}) = 0.208$ is the standard deviation. Such a Normal random variable is entered into cell H12 of Figure 8. If you watch cell H12 and cell E15 through many recalculations, the only difference in their pattern of behavior that you should observe is that the average in cell E15 is always a multiple of 1/30. (To show more precisely that the sample average in cell E15 has a probability distribution very close to that of the Normal random variable in cell H12, you could make a simulation table containing several hundred independently recalculated values of each of these random variables. Then by sorting each column in this simulation table, you could make a chart that estimates the inverse cumulative distribution of each random variable. These two curves should be very close.)

This central limit theorem is the reason why, of all the formulas that people could devise for measuring the center and the spread of probability distributions, the expected value and the standard deviation have been the most useful for statistics. Other probability distributions that have the same expected value 3 and standard deviation 1.14 could be quite different in other respects, but the central limit theorem tells us that an average of 30 independent samples from any such distribution would behave almost the same. (For example, try the probability distribution in which the possible values are 2, 3, and 7, with respective probabilities $P(2)=0.260$, $P(3)=0.675$, and $P(7)=0.065$.)

Now suppose that we did not know the expected value of \mathbf{K} , but we did know that its standard deviation was $\sigma=1.14$, and we knew how to simulate \mathbf{K} . Then we could look at any average of n independently simulated values and we could assign 95% probability to the event that our sample average does not differ from the true expected value by more than $1.96*\sigma/(n^{0.5})$. That is, if we let \mathbf{Y}_n denote the average of our n simulated values, then the interval from $\mathbf{Y}_n - 1.96*\sigma/(n^{0.5})$ to $\mathbf{Y}_n + 1.96*\sigma/(n^{0.5})$ would include the true $E(\mathbf{K})$ with probability 0.95. This interval is called a 95% confidence interval. With $n=30$ and $\sigma=1.14$, the radius r (that is, the distance from the center to either end) of this 95% confidence interval would be

$$r = 1.96*\sigma/(n^{0.5}) = 1.96*1.14/(30^{0.5}) = 0.408.$$

If we wanted the radius of our 95% confidence interval around the sample mean to be less than

some number r , then we would need to increase the size of our sample so that

$$1.96*\sigma/(n^{0.5}) < r, \text{ and so } n > (1.96*\sigma/r)^2$$

For example, to make the radius of our 95% confidence interval smaller than 0.05, the sample size n must be

$$n > (1.96*\sigma/r)^2 = (1.96*1.14/0.05)^2 = 1997.$$

Now consider the case where we know how to simulate an unknown quantity but we do not know how to calculate its expected value or its standard deviation. In this case, where our confidence-interval formula calls for the unknown probabilistic standard deviation σ , we must replace it by the sample standard deviation that we compute from our simulation data. If the average of n independent simulations is X and the sample standard deviation is S , then our estimated 95% confidence interval for the true expected value is from $X-1.96*S/(n^{0.5})$ to $X+1.96*S/(n^{0.5})$, where the quantity $S/(n^{0.5})$ is our estimated standard deviation of the sample average. In Figure 8, for example, the sample standard deviation S is computed in cell E16 by the formula `=STDEV(A15:C24)`, the sample size n is computed in cell D10 by the formula `=COUNT(A15:C24)`, the quantity $S/(n^{0.5})$ is computed in cell E17 by the formula

$$=E16/(D10^{0.5})$$

and then a 95% confidence interval for $E(\mathbf{K})$ is calculated in cells E20 and F20 by the formulas

$$=E15-1.96*E17 \text{ and } =E15+1.96*E17$$

(Recall that E15 is the sample average.)

To say that the interval from E20 to F20 is a 95% confidence interval for the true expected value is to say that the true expected value of 3 (in cell D7) should be between these two numbers 95% of the time when the spreadsheet in Figure 8 is recalculated many times independently. You can verify this by watching cell E22 while recalculating. Cell E22 contains the formula

$$=AND(E20<D7,D7<F20)$$

and so it should read TRUE about 95% of the time (about 19 times out of 20). (Here we use the Excel function `AND(statement1,statement2)`, which returns the value TRUE if statement1 and statement2 are both TRUE, and returns FALSE otherwise.)

So based only on the 30 independent simulations shown in Figure 8, the formulas in cells

E20 and F20 give us a 95% confidence interval from 2.715 to 3.550, that is 3.133 ± 0.417 , for $E(\mathbf{K})$. If the radius 0.417 seems too large, then we could get a narrower 95% confidence interval by using a larger sample. If we only knew this estimated sample standard deviation of 1.166, we could estimate that the radius of our 95% confidence interval could be reduced below 0.05 if the sample size n were increased so that $1.96 * 1.166 / (n^{0.5})$ were less than 0.05, which happens when n is greater than $(1.96 * 1.166 / 0.05)^2 = 2089$.

It is important to understand the difference between cells E16 and E17 in Figure 8. The value of cell E16 is the sample standard deviation $STDEV(A15:C24)$. So E16 is our statistical estimate of the standard deviation of any one random cell in the range A15:C24. The value of cell E17 is the sample standard deviation in E16 divided by the square root of the sample size. So E17 is our estimate of the standard deviation of the random cell E15, which is average of all the random cells in the range A15:C24. When we recalculate the 30 simulated values in cells A15:C24, the sample average tends to vary less than any one cell in the sample range, because when one cell is relatively high there is usually some other relatively low cell to cancel it out. That is why the standard deviation of the sample average is smaller than the standard deviation of any one cell in the sample, by a factor of $1/n^{0.5}$.

(You might wonder whether we should broaden our 95% confidence intervals when we use an estimated sample standard deviation instead of the true standard deviation. The answer is that we should, but the adjustments are relatively minor unless the sample size is small. The corrected formula is based on something that statisticians call a T-distribution. When the sample size n is small, we should replace the constant 1.96 in our 95% confidence formulas by the value of the Excel formula $TINV(0.05, n-1)$. When n is 30, this value is 2.045. As the sample size n increases, the value of $TINV(0.05, n-1)$ rapidly approaches 1.96, and so we will not worry about this T-adjustment in this course.)

A corollary of the central limit theorem can also be used to tell us something about the accuracy of our statistical estimates of points on the (inverse) cumulative probability curve. Suppose that \mathbf{X} is a random variable with a probability distribution that we know how to simulate, but we do not know how to directly compute its cumulative probability curve. For any number y , let $Q_n(y)$ denote the percentage of our sample that is less than y , when we get a sample

of n independent values drawn from the probability distribution of \mathbf{X} . Then we should use $Q_n(y)$ as our estimate of $P(\mathbf{X} < y)$. But how good is this estimate? When n is large, $Q_n(y)$ is a random variable with an approximately Normal distribution, its expected value is $P(\mathbf{X} < y)$, and its standard deviation is $(P(\mathbf{X} < y)(1 - P(\mathbf{X} < y)) / n)^{0.5}$. But this standard deviation is always less than $0.5 / (n^{0.5})$. Of course $1.96 * 0.5 / (n^{0.5})$ is only slightly less than $1 / n^{0.5}$. So around any point $(Q_n(y), y)$ in an estimated inverse cumulative probability curve like Figure 7, we could put a horizontal confidence interval over the cumulative probabilities from $Q_n(y) - 1 / n^{0.5}$ to $Q_n(y) + 1 / n^{0.5}$, and this interval would have a probability greater than 95% of including the true cumulative probability at y . When n is 400, for example, the radius of this cumulative-probability interval around $Q_n(y)$ is $1 / n^{0.5} = 1 / 20 = 0.05$. If we wanted to reduce the radius of this 95%-confidence interval below 0.02, then we would increase the sample size to $1 / 0.02^2 = 2500$.

7. Decision criteria

We have used the "Superior Semiconductors (A)" case as an example to introduce basic ideas of probability and statistics. But after all these ideas have been introduced, we are still left with a decision problem. On the basis of our analysis of the uncertainty in this situation, should we recommend that Superior Semiconductors introduce the new T-regulator product or not? To answer this question, we need some fundamental assumption about what determines an optimal decision under uncertainty. The assumption that we will usually apply in this course is the criterion of expected value maximization (or the expected value criterion).

Any quantitative decision analysis must involve some numerical measure of payoff, such that increasing the decision-maker's payoff is always considered an improvement. For most economic decision problems, net monetary returns or profit (or the present-discounted value of future profit) may be identified as the "payoff" that the decision-maker wants to increase whenever possible. But when there is uncertainty, we do not know whether a particular decision (like that of introducing the new T-regulator product) would increase or decrease the decision-maker's payoff. The criterion of expected value maximization asserts that, among the various alternatives that are available to a decision-maker, the optimal decision is the one that yields the highest expected value of the decision-maker's payoff.

So if we take Superior Semiconductor's profit to be the measure of "payoff" in this decision problem, then the optimal decision can be identified simply by computing the expected value of Superior Semiconductor's profit from the proposed new product. As we have seen, this expected value is

$$E(\text{Profit}) = 0.10*24 + 0.25*7.33 + 0.30*(-1) + 0.25*(-6) + 0.10*(-9.33) = 1.5$$

The alternative of not introducing the new product generates an expected profit of 0, of course. So by the criterion of expected value maximization, Superior Semiconductors should introduce the new product, because $1.5 > 0$.

The expected value formula has many good properties to recommend it as a criterion for decision-making under uncertainty. It takes account of all possible outcomes in a sensible way, and it is more sensitive to possible outcomes that are more likely. The argument for expected value maximization is particularly compelling in games that can be repeated. If we know that we will face a given type of decision problem many times, but with the new outcome being determined independently each time, then a strategy of choosing the alternative that yields the highest expected value will almost surely maximize our long-term total payoff, by the law of large numbers.

This expected-value criterion may be interpreted to mean that all we should care about is the expected value of some appropriately measured payoff. But this interpretation can lead to trouble if the words "of ... payoff" are forgotten. If you thought that our expected-value criterion meant that we should only care about the expected number of competitors, then you would act as though the number of competitors would be 3, in which case profit would be -1 , and your recommendation would be to not introduce the new product. The error here is to compute the expected value of the wrong random variable (not payoff), and then to try to compute an expected payoff from it with the "flaw of averages" (which we discussed in section 4).

	A	B	C	D	E	F	G	H	
1	"SUPERIOR SEMICONDUCTOR (A)"				26	FixedCost (\$millions)			
2					100	MarketValue (\$millions)			
3	Let K = (unknown number of competitors entering market)								
4	Probability distribution of K								
5		k	P(K=k)		Profit (\$millions) if k=#competitors				
6		1	0.10		24.00				
7		2	0.25		7.33				
8		3	0.30		-1.00				
9		4	0.25		-6.00				
10		5	0.10		-9.33				
11									
12	Computations from the probability distribution								
13		E(K)			E(Profit)				
14		3			1.5				
15									
16	Simulation model:								
17		K			Profit				
18		1			24.00				
19									
20	Statistics from simulations:								
21		1.497922	E(Profit) est'd from SimTable						
22		9.281945	Stdev(Profit) est'd from SimTable						
23		-9.33333	Value at risk (percentile =				0.05)		
24		401	Sample size						
25			95% confidence interval for E(Profit)						
26		Sim'd profit			1.034404	1.96144			
27	SimTable	24.00							
28	0	-1.00	FORMULAS FROM RANGE A1:G28						
29	0.0025	-1.00	E6. = $\$E\$2/(1+B6)-\$E\1						
30	0.005	-6.00	E6 copied to E6:E10						
31	0.0075	7.33	B14. =SUMPRODUCT(B6:B10,\$C\$6:\$C\$10)						
32	0.01	-9.33	E14. =SUMPRODUCT(E6:E10,\$C\$6:\$C\$10)						
33	0.0125	24.00	B18. =DISCRINV(RAND(),B6:B10,C6:C10)						
34	0.015	-9.33	E18. = $\$E\$2/(1+B18)-\$E\1						
35	0.0175	7.33	B27. =E18						
36	0.02	-1.00	B21. =AVERAGE(B28:B428)						
37	0.0225	-1.00	B22. =STDEV(B28:B428)						
38	0.025	7.33	B23. =PERCENTILE(B28:B428,G23)						
39	0.0275	-6.00	B24. =COUNT(B28:B428)						
40	0.03	-1.00	E26. =B21-B22/(B24^0.5)						
41	0.0325	-6.00	F26. =B21+B22/(B24^0.5)						

Figure 9. Analysis of "Superior Semiconductor (A)" case.

Figure 9 shows a decision analysis of the "Superior Semiconductor (A)" case. In cell E13, the expected value is profit is calculated directly from the probability distribution, and this number

is really all we need to recommend the new product under the expected value criterion. But to illustrate what we would do if we could not compute expected profit directly from the probability distribution, the spreadsheet also contains a table of 401 independent simulations of the unknown profit, and the average of these profits is exhibited in cell B21 as an estimator of the expected profit. Even if we could not see the true expected value in cell E13, our simulation data is strong enough to support reasonable confidence that the expected value is greater than 0, because we find a positive lower bound (1.034404) in the 95% confidence interval for expected profit that is computed in cells E26 and F26.

But now, having advocated the expected value criterion, I now have to admit that it is often not fully satisfactory as a basis for decision-making. In practice, people often prefer decision alternatives that yield lower expected profits, when the alternatives that yield higher expected profits are also more risky. So in any serious decision analysis, we should go beyond simply reporting expected payoff values, and we should also report some measures of the risks associated with the alternatives that are being considering.

As we have seen, the standard deviation is often used as a measure of the spread of likely outcomes of an unknown quantity, and so the standard deviation of profit may be used as a measure of risk. Thus, cell B22 in Figure 9 estimates the standard deviation of profit from the proposed new product in this case. The large size of this sample standard deviation (9.28 \$million, much larger than the expected value) is a strong indication that this new product should be seen as a very risky.

Another measure of risk that has gained popularity in recent years is called "value at risk," often abbreviated "VaR" (note the capitalization of outer letters only). The value at risk is defined to level of net profit that has some pre-specified cumulative probability, usually either 0.05 or 0.10. Cell B23 computes the value at risk for the 5% cumulative-probability level from our simulation data, using the formula =PERCENTILE(B28:B428,G23), where cell G23 contains the cumulative-probability value 0.05 and cells B28:B428 contains our simulated profit data.

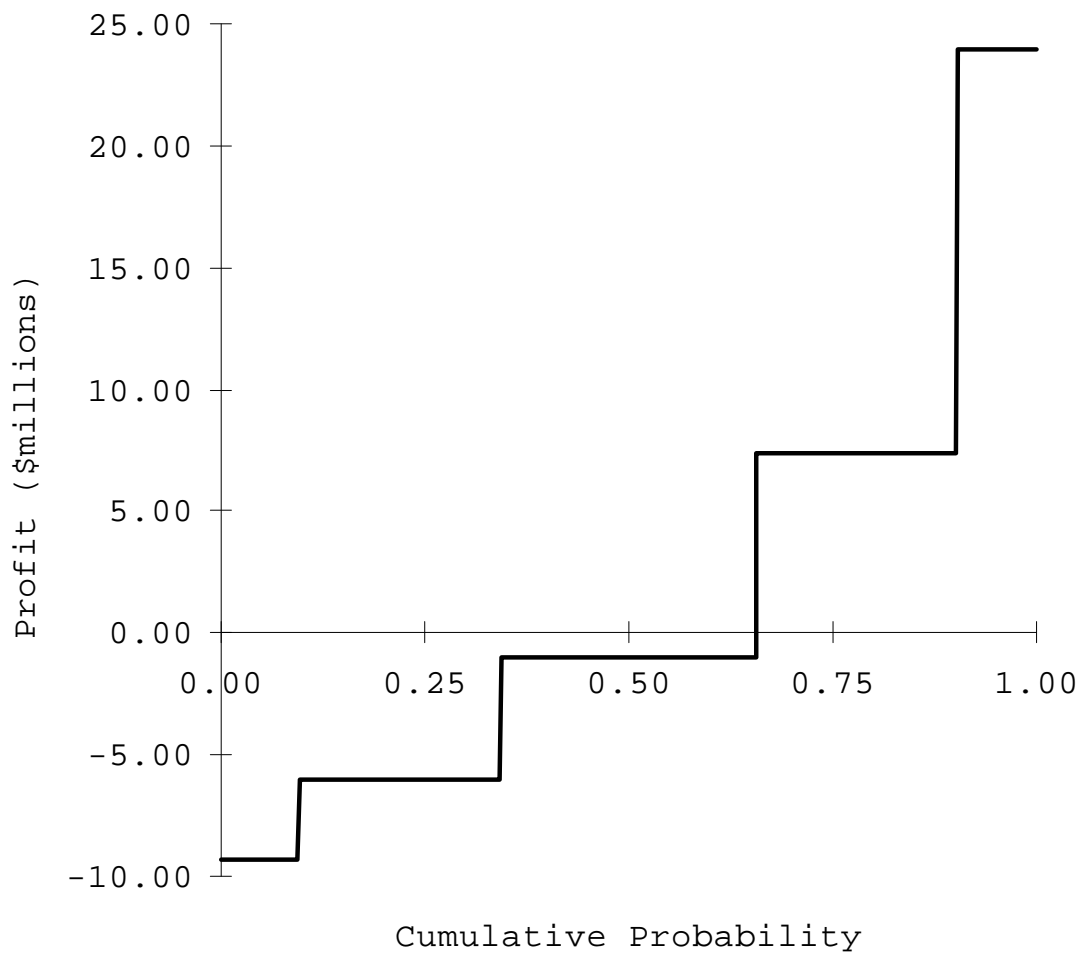


Figure 10. Cumulative risk profile for "Superior Semiconductors (A)" case.

The cumulative risk profile for a decision may be defined as the inverse cumulative probability distribution of the payoffs that would result from this decision. Figure 10 shows the cumulative risk profile for the decision to introduce the new product in this case. This cumulative risk profile was made from the simulation table in Figure 9. First the simulated profit data in cells B28:B428 were sorted, and then these sorted profit values were plotted on the vertical axis of an XY-chart, with the percentile index in cells A28:A428 plotted on the horizontal axis. Notice that this cumulative risk profile contains all the information about the value at risk, for any probability

level. By definition, the value at risk for the cumulative-probability level 0.05 is just the height of the cumulative risk profile at 0.05 on the horizontal cumulative-probability axis. So the cumulative risk profile may give the most complete overall picture of the risks associated with a decision.

The limitations of the expected value maximization as a criterion for decision-making have been addressed by two important theories of economic decision-making: utility theory and arbitrage-pricing theory. Utility theory is about decision-making by individuals, whereas arbitrage-pricing theory tells us about decision-making for publicly held corporations. These theories will be introduced near the end of this course, but I can give a brief description of each of these theories now.

Utility theory shows how to apply an expected value criterion to any individual decision-maker who satisfies some intuitive consistency assumptions. The main result of utility theory is that any rational decision-maker should have some way of measuring payoff, called a utility function, such that the decision-maker will always prefer the alternative that maximizes his or her expected utility. So if a risk-averse decision-maker does not want to make all decisions purely on the basis of his expected monetary payoff, that only means that his utility payoff is different from his monetary payoff. Utility theory then teaches us how to assess an individual's risk tolerance and derive a utility function that gives us a utility value corresponding to any monetary value of profit or income. So the good news is that, once we learn how to assess utility functions for risk-averse decision-makers, the mathematical method that we will use to analyze any decision problem will still be expected value maximization. The only difference is that, instead of maximizing an expected monetary value, we will maximize an expected utility value.

But consider now the situation where a corporate officer makes decisions on behalf of the stockholders who own a publicly held corporation. There may be millions of stockholders, and each of them may have a different utility function for profit. What principles of optimal decision-making can we recommend for such a situation? One approach is to assume that all stockholders want to maximize the value of the firm, understanding that a financial market can assign a value to any uncertain stream of future income. So a question about whether some new project is worth undertaking becomes a question about whether the financial market would assign a value to the

uncertain profits of this project that is greater than its current cost. Arbitrage pricing theory then tells us that coherent financial markets should evaluate such uncertain prospects according to a criterion of maximizing expected present-discounted value of profits, except that the probabilities that we use in these expected-value calculations may be some market-adjusted probabilities which differ from the relative-frequency probabilities that statisticians use. Typically, an event may have a market-adjusted probability that is lower than its statistical frequency if it is an event in which the overall stock market goes up, whereas the market-adjusted probability may be higher than the statistical frequency for an event in which the overall stock market goes down. These adjustments are occur because a risk-averse investor may prefer to take an actuarially unfair bet that shifts future income from events where his other investments will do well, to events where his other investments will do badly. But again our good news is that sophisticated financial analysis of corporate decisions is compatible with a criterion of expected value maximization. The only difference is that, in these expected value calculations, the probabilities of some events may need to be adjusted to take account of how the overall performance of financial markets may depend on these events.

8. Summary

In this chapter, we focused on a simple decision problem involving a single unknown quantity that has only finitely many possible values. In this context, we introduced some basic concepts for describing discrete probability distributions: the expected value or mean of the distribution, the standard deviation and the variance, and cumulative probability charts. We saw how to make a random variable with any given probability distribution by using the inverse cumulative-probability function with a `RAND()` as input. The Simtools function `DISCRINV` was introduced to facilitate such simulations.

We then introduced techniques for estimating expected values, standard deviations, and cumulative probabilities from simulation data, using the law of large numbers for assurance that these estimates are very likely to be quite accurate if the sample size is very large. For a more precise assessment of the accuracy of the sample average as an estimate for an unknown expected value, we introduced Normal distributions and the central limit theorem. We learned that a

sample average, as a random variable, has a standard deviation that is inversely proportional to the square root of the sample size. We then saw how to compute a 95% confidence interval for the expected value of a random variable, using simulation data.

Finally, the criterion of expected value maximization was discussed. So the expected value of a suitably-measured payoff quantity was recommended as the best single number to guide decision-making under uncertainty. The standard deviation of payoff, the value at risk (for some pre-specified cumulative-probability level), and the entire cumulative risk profile were recommended as also worth reporting in a decision analysis, to better describe the levels of risk entailed by different decision alternatives.

Excel functions used in this chapter include AND, AVERAGE, NORMINV, STDEV, and SUMPRODUCT. Simtools functions introduced in this chapter include DISCRINV and STDEVPR. We also used the Data>Sort and Insert>Chart>XY(Scatter) commands to make inverse cumulative charts from simulation data.

Exercises

1. Let \mathbf{X} denote an unknown quantity that has three possible values: 2, 3, and 7, and suppose that their probabilities are $P(\mathbf{X}=2) = 0.260$, $P(\mathbf{X}=3) = 0.675$, $P(\mathbf{X}=7) = 0.065$.

Let \mathbf{Y} denote another unknown quantity that has three possible values: -1, 3, and 4, and suppose that their probabilities are $P(\mathbf{Y}=-1) = 0.065$, $P(\mathbf{Y}=3) = 0.675$, $P(\mathbf{Y}=4) = 0.260$.

(a) Compute $E(\mathbf{X})$, $\text{Stdev}(\mathbf{X})$, $E(\mathbf{Y})$ and $\text{Stdev}(\mathbf{Y})$.

(b) According to the central limit theorem, an average of 36 random variables drawn from the probability distribution of \mathbf{X} should have approximately what probability distribution? (Be sure to specify the mean and standard deviation.)

(c) In a spreadsheet, make a simulation table that tabulates values of five random variables as follows:

the first is a single cell that simulates \mathbf{X} ,

the second is a single cell that simulates \mathbf{Y} ,

the third is an average of 36 cells independently drawn from the probability distribution of \mathbf{X} ,

the fourth is an average of 36 cells independently drawn from the probability distribution of \mathbf{Y} ,

the fifth is a single random cell drawn from the probability distribution that you predicted in (b).

Include at least 400 data rows in your simulation table. (This calculation may take a few minutes on older computers.)

(d) Using your simulation table in (c), compute the sample mean and standard deviation for each of the five random variables, and make an XY-chart that estimates the (inverse) cumulative distribution for these five random variables. (*Hint on charting keystrokes: You can separately sort each of your five columns of simulation data, then select the percentile index and five sorted data columns in the simulation table, and insert an XY-chart.*)

2. In a simulation table with data from 400 independent simulations of a random variable \mathbf{W} , the sample mean is 220.12, and the sample standard deviation is 191.63.

(a) Estimate the standard deviation of the sample mean when the sample size is 400.

(b) Based on this data, compute a 95% confidence interval for the true expected value of this random variable $E(\mathbf{W})$.

(c) Suppose that we want to make a new table of simulation data which will generate a 95% confidence interval for $E(\mathbf{W})$ that has a radius of about 5. How large should this new simulation table be? (That is, how many independent simulations should it include?)

3. What is the discrete probability distribution of the random variable that would be generated by each of the following Excel formulas? Check your answer by a large simulation.

(a) =IF(RAND() $>$ 0.3,2,0)+IF(RAND() $>$ 0.6,3,0)

(b) =IF(AND(0.3<RAND(),RAND() $<$ 0.4),1,0)

(c) =IF(RAND() $<$ 0.6,IF(RAND() $<$ 0.5,1,2),3)

(d) How would your answers change if we entered =RAND() into cell A1 and we replaced every RAND() in the above formulas by a reference to cell A1?