

Kellogg Graduate School of Management  
Practical Probability with Spreadsheets

Chapter 1: INTRODUCTION TO PROBABILITY AND SIMULATION IN SPREADSHEETS

Probability is the basic mathematics of uncertainty. Whenever there is something that we do not know, our uncertainty can (in principle) be described by probabilities. This document is meant to be an elementary introduction to the basic ideas of probability. My goal is to do more than simply introduce these ideas, however. Right from the start, I will try to show you how to use the ideas of probability to build simulation models in spreadsheets.

Everything that I do in spreadsheets will be described as I do it with Microsoft Excel 5.0 or higher (in Windows). Within Excel, there are many different ways of telling the program to do any one task. Most commands can begin by selecting one of the options on the menu bar at the top of the screen ("File Edit View Insert Format Tools Data Window Help"), and you can make this selection either by clicking with the mouse or by pressing the [Alt] key and the underlined letter key. Then, secondary options appear under the menu bar, and you select among them by either clicking with the mouse or by typing the underlined letter key. Many common command sequences can also be entered by a short-cut keystroke (which is indicated in the pop-down menus), or by clicking on a button in a power bar that you can display on the screen (try the View menu). In this essay, I will describe command descriptions as if you always use the full command sequence from the top menu.

I will also assume that you have an add-in for Excel called **simtools.xla** which you can download from the Internet at

**<http://www.kellogg.nwu.edu/faculty/myerson/ftp/addins.htm>**

When you download simtools.xla, you should save it in your macro library subdirectory under the Excel directory of your computer's hard disk. Then, in Excel, you can install Simtools by using the Tools>AddIns command sequence and then selecting the "Simulation Tools" option that will appear in the dialogue box. Once installed, "Simtools" should appear as an option directly in Excel's Tools menu.

## 1. How to toss coins in a spreadsheet

When we study probability theory, we are studying uncertainty. To study uncertainty with a spreadsheet, it is useful to create some uncertainty within the spreadsheet itself. Knowing this, the designers of Excel gave us one simple but versatile way to create such uncertainty: the RAND() function. I will now describe how to use this function to create a spreadsheet that simulates tossing coins (a favorite first example of probability teachers).

With the cursor on cell A1 in the spreadsheet, let us type the formula

```
=RAND ( )
```

and then press the Enter key. A number between 0 and 1 is displayed in cell A1. Then (by mouse or arrow keys) let us move the cursor to cell B1 and enter the formula

```
=IF(A1<0.5, "Heads", "Tails")
```

(The initial equals sign [=] alerts Excel to the fact that what follows is to be interpreted as a mathematical formula, not as a text label. The value of an IF(•,•,•) function is its second parameter if the first parameter is a true statement, but its value is the third parameter if the first is false.) Now Excel checks the numerical value of cell A1, and if A1 is less than 0.5 then Excel displays the text "Heads" in cell B1, but if A1 is greater than or equal to 0.5 then Excel displays the text "Tails" in cell B1.

If you observed this construction carefully, you would have noticed that the number in cell A1 changed when the formula was entered into cell B1. In fact, every time we enter anything into spreadsheet, Excel recalculates the everything in the spreadsheet and it picks a new value for our RAND() function. (I am assuming here that Excel's calculation option is set to "Automatic" on your computer. If not, this setting can be changed under Excel's Tools>Options menu.) We can also force such recalculation of the spreadsheet by pushing the "Recalc" button, which is the [F9] key in Excel. If you have set up this spreadsheet as we described above, try pressing [F9] a few times, and watch how the number in cell A1 and the text in cell B1 change each time.

Now let us take hands away from the keyboard and ask the question: What will be the next value to appear in cell A1 after the next time that the [F9] key is pressed? The answer is that we do not know. The way that the Excel program determines the value of the RAND() function each time is, and should remain, a mystery to us. (The parentheses at the end of the function's

name may be taken as a sign that the value depends on things that we cannot see.) My vague understanding is that the program does some very complicated calculations, which may depend in some way on the number of seconds past midnight on the computer's clock, but which always generate a value between 0 and 1. I know nothing else specific about these calculations. The only thing that you and I need to know about these RAND() calculations is that the value, rounded to any number of decimal digits, is equally likely to be any number between 0 and 1. That is, the first digit of the decimal expansion of this number is equally likely to be 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. Similarly, regardless of the first digit, the second decimal place is equally likely to be any of these digits from 0 to 9, and so on. Thus, the value of RAND() is just as likely to be between 0 and .1 as it is to be between .3 and .4. More generally, for any number  $v$ ,  $w$ ,  $x$ , and  $y$  that are between 0 and 1, if  $v - w = x - y$  then the value of the RAND() expression is as likely to be between  $w$  and  $v$  as it is to be between  $y$  and  $x$ . This information can be summarized by saying that, from our point of view, RAND() is drawn from a uniform probability distribution over the interval from 0 to 1.

The cell B1 displays "Heads" if the value of A1 is between 0 and 0.5, whereas it displays "Tails" if A1 is between 0.5 and 1. Because these two intervals have the same length ( $0.5 - 0 = 1 - 0.5$ ), these two events are equally likely. That is, based on our current information, we should think that, after we next press [F9], the next value of cell B1 is equally likely to be Heads or Tails. So we have created a spreadsheet cell that behaves like a fair coin toss every time we press [F9]. We have our first simulation model.

After pressing [F9] to relieve your curiosity, you should press it a few more times to verify that, although it is impossible to predict whether Heads or Tails will occur next in cell B1, they tend to happen about equally often when we recalculate many times. It would be easier to appreciate this fact if we could see many of these simulated coin tosses at once. This is easy to do by using the spreadsheet's Edit>Copy and Edit>Paste commands to make copies of our formulas in the cells A1 and B1. So let us make copies of this range A1:B1 in all of the first 20 rows of the spreadsheet. (Any range in a spreadsheet can be denoted by listing its top-left and bottom right cells, separated by a colon.)

To copy in Excel, we must first select the range that we want to copy. This can be done

by moving the cursor to cell A1 and then holding down the shift key while we move the cursor to B1 with the right arrow key. (Pressing an arrow key while holding down the shift key selects a rectangular range that has one corner at the cell where the cursor was when the shift key was first depressed, and has its opposite corner at the current cell. The selected range will be highlighted in the spreadsheet.) Then, having selected the range A1:B1 in the spreadsheet, open the Edit menu and choose Copy. Faint dots around the A1:B1 range indicate that this range has been copied to Excel's "clipboard" and is available to be pasted elsewhere. Next, select the range A1:A20 in the spreadsheet, and then open the Edit menu again and choose Paste. Now the spreadsheet should look something like Figure 1:

	A	B	C	D	E	F	G	H
1	0.637032	Tails						
2	0.823076	Tails						
3	0.877162	Tails						
4	0.503784	Tails						
5	0.204999	Heads						
6	0.333061	Heads						
7	0.163438	Heads						
8	0.514505	Tails						
9	0.332584	Heads						
10	0.065892	Heads						
11	0.452892	Heads						
12	0.833748	Tails						
13	0.252074	Heads						
14	0.021183	Heads						
15	0.951513	Tails						
16	0.645427	Tails						
17	0.725836	Tails			FORMULAS FROM RANGE A1:D20			
18	0.990074	Tails			A1.	=RAND()		
19	0.46103	Heads			B1.	=IF(A1<0.5,"Heads","Tails")		
20	0.285434	Heads			A1:B1 copied to A2:A20			

**Figure 1. Simple coin-tossing spreadsheet.**

(The descriptions that appear in the lower right corner of Figure 1 are just text that I typed into cells E17:E20, with the help of the add-in *formlist.xls*.)

In Figure 1, we have made twenty copies of the horizontal range A1:B1, putting the left-hand side of each copy in one of the cells in the vertical range A1:A20. So each of these

twenty A-cells contains the RAND() function, but the values that are displayed in cells A1:A20 are different. The value of each RAND() is calculated independently of all the other RANDs in the spreadsheet. The spreadsheet even calculates different RANDs within one cell independently, and so a cell containing the formula =RAND() - RAND() could take any value between -1 and +1.

The word "independently" is being used here in a specific technical sense that is very important in probability theory. When we say that a collection of unknown quantities are independent of each other, we mean that learning the values of some of these quantities would not change our beliefs about the other unknown quantities in this collection. So when we say that the RAND() in cell A20 is independent of the other RANDs in cells A1:A19, we mean that knowing the values of cells A1:A19 tells us nothing at all about the value of cell A20. If you covered up cell A20 but studied the values of cells A1:A19 very carefully, you should still think that the value of cell A20 is drawn from a uniform distribution over the interval from 0 to 1 (and so, for example, is equally likely to be above or below 0.5), just as you would have thought before looking at any of these cell values.

Each of the twenty cells in B1:B20 contains a copy of the IF... function that we originally entered into cell B1. If you run the cursor through the B-cells, however, you should notice that the reference to cell A1 in cell B1 was adjusted when it was copied. For example, B20 contains the formula

=IF(A20<.5, "Heads", "Tails")

Excel's Copy command treats references to other cells as relative, unless we preface them with dollar signs (\$) to make them absolute. So each of the copied IF functions looks to the cell to the left for the number that it compares to .5, to determine whether "Heads" or "Tails" is displayed. Thus we have set up a spreadsheet in which cells B1:B20 simulate twenty independent tosses of fair coins.

Now let us change this spreadsheet so that can do something that you could not do so easily with coins: simulate twenty independent tosses of an unfair coin that is not equally likely to be Heads or Tails. Into cell B1, enter the formula

=IF(A1<\$D\$1, "Heads", "Tails")

(You can use the edit key [F2] to get into the old formula and revise it, simply changing the 0.5 to \$D\$1.) Next, copy cell B1 and paste it to B2:B20. The dollar signs in the formula tell Excel to treat the reference to D1 as an absolute, not to be adjusted when copied, and so the formula in cell B20 (for example) should now be =IF(A20<\$D\$1,"Heads","Tails"). Now enter any number between 0 and 1 into cell D1. If you enter 0.25 into cell D1, for example, then your spreadsheet may look something like Figure 2.

	A	B	C	D	E	F	G	H
1	0.760253	Tails		0.25				
2	0.923893	Tails						
3	0.192721	Heads						
4	0.502798	Tails						
5	0.566062	Tails						
6	0.518313	Tails						
7	0.545495	Tails						
8	0.845976	Tails						
9	0.726467	Tails						
10	0.807506	Tails						
11	0.13915	Heads						
12	0.162705	Heads						
13	0.372462	Tails						
14	0.509231	Tails						
15	0.281672	Tails						
16	0.341317	Tails						
17	0.171075	Heads			FORMULAS FROM RANGE A1:D20			
18	0.519393	Tails			A1. =RAND()			
19	0.51953	Tails			B1. =IF(A1<\$D\$1,"Heads","Tails")			
20	0.449476	Tails			A1:B1 copied to A2:A20			

**Figure 2. Coin-tossing with adjustable probabilities.**

In this spreadsheet, each cell in B1:B20 will display "Heads" only if the random number to the left is between 0 and 0.25; but it is three times more likely that the number will be above 0.25, and so we should expect substantially more Tails than Heads. Introducing more language of probability theory, we may say that, in each cell in the range B1:B20 in this spreadsheet, the probability of getting "Heads" after the next recalculation is 0.25. If we entered any other number  $p$  into cell D1, then this probability of getting a Heads in each B-cell would change to this new probability  $p$ , independently of the other B-cells.

More generally, when I say that the probability of some event "A" is some number  $q$  between 0 and 1, I mean that, given my current information, I think that this event "A" is just as likely to occur as the event that any single `RAND()` in a spreadsheet will take a value less than  $q$  after the next recalculation. That is, I would be indifferent between a lottery ticket that would pay me \$100 if the event "A" occurs and a lottery ticket that would pay me \$100 if the `RAND()`'s next value is less than the number  $q$ . When this is true, I may write the equation  $P(A) = q$ .

## 2. A simulation model of twenty sales calls

Perhaps it seems somewhat frivolous to model 20 coin tosses. So let us consider instead a salesperson who will call on 20 customers this week. In each sales call, the salesperson may make a sale or not. To adapt the spreadsheet in Figure 2 to this situation, let us begin by inserting a new row at the top of the spreadsheet, into which we write labels that describe the interpretation of the cells below. We can insert a new row 1 by selecting cell A1 and entering the command sequence `Insert>Rows`. Notice that formula references to cell `$D$1` automatically change to cell `$D$2` when we insert a new row 1 at the top.

Let us enter the label 'Sales:' in the new empty cell B1. Below, in the 20 cells of range B2:B21, we will simulate the outcomes of the calls to the 20 customers, with a 1 denoting a sale from this call and 0 denoting no sale from this call. So let us re-edit cell B2, entering the formula

$$=IF(A1<D2,1,0)$$

into cell B2. Recall that we also have `=RAND()` in cell A2. To model the other 19 calls, let us copy A2:B2 and paste it to A3:A21. If we enter the number 0.5 into cell D2 then, with our new interpretation, this spreadsheet simulates a situation in which the salesperson has a probability 1/2 of making a sale with each of his 20 customers. So we should enter the label

'P(Sale) in each call' into cell D1. To indicate that the values in column A are just random numbers that we use in building our simulation model, we may enter the label

'(rands)' in cell A1.

Indicating sales and no-sales in the various calls by 1s and 0s makes it easy for us to count the total number of sales in the spreadsheet. We can simply enter the formula `=SUM(B2:B21)`

into cell D10 (say), and enter the label 'Total Sales' into the cell above. The result should look similar to Figure 3.

	A	B	C	D	E	F	G
1	(rands)	Sales:		P(Sale in each call)			
2	0.713543	0		0.5			
3	0.914063	0					
4	0.223945	1					
5	0.570389	0					
6	0.000528	1					
7	0.396419	1					
8	0.445235	1					
9	0.861853	0		Total Sales			
10	0.474134	1		13			
11	0.680675	0					
12	0.116178	1					
13	0.194609	1					
14	0.469929	1					
15	0.385216	1					
16	0.419566	1					
17	0.772312	0			FORMULAS FROM RANGE A1:D21		
18	0.08889	1			A2. =RAND( )		
19	0.197045	1			B2. =IF(A2<D2,1,0)		
20	0.392369	1			A2:B2 copied to A3:A21		
21	0.664038	0			D10. =SUM(B2:B21)		

**Figure 3. Simple model of independent sales calls.**

This Figure 3 simulates the situation for a salesperson whose selling skill is such that he has a probability 0.50 of making a sale in any call to a customer like these twenty customers. A more skilled salesperson might be more likely to make a sale in each call. To simulate 20 calls by a more (or less) skilled salesperson, we could simply replace the 0.50 in cell D2 by a higher (or lower) number that appropriately represents the probability that this salesperson will get a sale from any such call, given his actual level of selling skill. In fact, we may think of the probability of making a sale in any call to a customer like these as being a numerical measure of the salesperson's "skill" at this kind of marketing.

But once the number 0.50 is entered into cell D2 in Figure 3, the outcomes of the 20 simulated sales calls are determined independently, because each depends on a different RAND variable in the spreadsheet. In this spreadsheet, if we knew that the salesperson made sales to all

of the first 19 customers, we would still think that his probability of making a sale to the 20th customer is  $1/2$  (as likely as a `RAND()` being less than 0.50). Such a strong independence assumption may seem very unrealistic. In real life, even if we knew that this company's salespeople generally make sales in about half of their calls, a string of 19 successful visits might cause us to infer that this particular salesperson is very highly skilled, and so we might think that he would be much more likely to get a sale on his 20th visit. On the other hand, if we learned that this salesperson had a string of 19 unsuccessful calls, then we might infer that he was probably unskilled, and so we might think him unlikely to make a sale on his twentieth call. To take account of such dependencies, we need to revise our model to one in which the outcomes of the 20 sales calls are not completely independent.

This independence problem is important to consider whenever we make simulation models in spreadsheets. The `RAND` function makes it relatively easy for us to make many random variables and events that are all independent of each other. Making random variables that are not completely independent of each other is more difficult. In Excel, if we want to make two cells not be independent of each other, then there must be at least one cell with a `RAND` function in it which directly or indirectly influences both of these cells. (You can trace all the cells which directly and indirectly influence any cell by repeatedly using the `Tools>Auditing>TracePrecedents` command sequence until it adds no more arrows. Then use `Tools>Auditing>RemoveAllArrows`.) So to avoid assuming that the sales simulated in B2:B21 are completely independent, we should think about some unknown quantities or factors that might influence all these sales events, and we should revise our spreadsheet to take account of our uncertainty about such factors.

Notice that the our concern about assuming independence of the 20 sales calls was really motivated in the previous discussion by our uncertainty about the salesperson's skill level. This observation suggests that we should revise the model so that it includes some explicit representation of our uncertainty about the salesperson's skill level. The way to represent our uncertainty about the salesperson's skill level is to make the skill level in cell D2 into a random variable. When D2 is random, then the spreadsheet will indeed have a random factor that influences all the twenty sales events.

To keep things as simple as possible in this introductory example, let us suppose (for now)



model is simulating a skilled salesperson. When the value is 0.333333, the spreadsheet model is simulating an unskilled salesperson. When the salesperson is skilled, he usually succeeds in more than half of his sales opportunities; and when he is unskilled, he usually fails in more than half of the calls. But if you recalculate this spreadsheet many times, you will occasionally see it simulating a salesperson who is skilled but who nevertheless fails in more than half of his calls.

Let us now ask a question that might arise in the work of a supervisor of such salespeople. If the salesperson sold to exactly 9 of the 20 customers on whom he called this week, then what should we think is the probability that he is actually skilled (but just had bad luck this week)? Remember: we are assuming that we believed him equally likely to be skilled or unskilled at the beginning of the week; but observing 9 sales in 20 gives us some information that should cause our beliefs to change.

To answer this question with our simulation model, we should recalculate the simulation many times. Then we can see how often do skilled salespeople make only 9 sales, and how often do unskilled salespeople make 9 sales. The relative frequencies of these two events in many recalculated simulations will give us a way to estimate how much can be inferred from the evidence of only 9 sales.

There is a problem, however. Our model is two-dimensional (spanning many rows and several columns), and it is not so easy to make hundreds of copies of it. Furthermore, even if we did put the whole model into one row of the spreadsheet, recalculating and storing hundreds of copies of 42 different numerical cells could strain the speed and memory of a computer.

But we do not really need a spreadsheet to hold hundreds of copies of our model. In our simulation, we only care about two things: is the salesperson skilled; and how many sales did he make? So if we ask Excel to recalculate our model many times, then we only need it to make a table that records the answers to these two questions for each recalculated simulation. All the other information (about which customers among the twenty in B2:B21 actually bought, and about the specific random values shown in A2:A21) that is generated in the repeated recalculations of the model can be erased as the next recalculation is done. Excel has the capability to make just such a table, which is called a "data table" in the language of spreadsheets. Simtools gives us a special version of the data table, called a "simulation table," which is

particularly convenient for analyzing such models.

To make a simulation table, the output that we want to tabulate from our model must be listed together in a single row. This model-output row must also have at least one blank cell to its left, and beneath the model-output row there must be many rows of blank cells where the simulation data will be written. In our case, the simulation output that we want is in cells D2 and D10, but we can easily repeat the information from cells in an appropriate model-output row. So to keep track of whether the simulated salesperson's skill is high, let us enter the formula `=IF(D2=2/3,1,0)` into cell B35, and to keep track of the number of sales achieved by this simulated salesperson, let us enter the formula `=D10` into cell C35. To remind ourselves of the interpretations of these cells, let us enter the labels 'Skill hi?' and '#Sales' into cells B34 and C34 respectively.

Now we select the range in which the simulation table will be generated. The top row of the selected range must include the model-output range B35:C35 and one additional unused cell to the left (cell A35). So we begin by selecting the cell A35 as the top-left cell of our simulation table. With the [Shift] key held down, we can then press the right-arrow key twice, to select the range A35:C35. Now, when we extend the selected range downwards, any lower row that we include in the selection will be filled with simulation data. If we want to record the results of about 1000 simulations, then we should include in our simulation table about 1000 rows below A35:C35. So continuing to hold down the [Shift] key, we can tap the [PgDn] key to expand our selection downwards. Notice that the number of rows and columns in the range size is indicated in the formula bar at the top left of the screen while we are using the [PgDn] or arrow keys with the [Shift] key held down. Let us expand the selection until the range A35:C1036 is selected, which gives us 1002 rows that in our selected range: one row at the top for model output and 1001 rows underneath it in which to store data. (We will see that SimTable output looks a bit nicer when the number of data rows is one more than a multiple of 100.) Finally, with this range A35:C1036 selected, we enter the command sequence

`Tools>Simtools>SimulationTable.`

After a pause for some computations, the result should look similar to Figure 5.

	A	B	C
30	FORMULAS		
31	B35. =IF(D2=2/3,1,0)		
32	C35. =D10		
33			
34	Skill hi?	#Sales	
35	SimTable	1	12
36	0	0	5
37	0.001	0	9
38	0.002	0	4
39	0.003	0	6
40	0.004	1	16
41	0.005	1	12
42	0.006	0	6
43	0.007	0	6
44	0.008	0	3
45	0.009	0	7

(Data continues to row 1036)

**Figure 5. Simulation table for model of twenty sales calls.**

What has happened? First, the simulation-table procedure wrote the word "Simtable" into cell A35, which was the top-left cell of the range that we selected. Then a border line was drawn under this cell and the model-output cells in row 35, to separate them visually from the cells below. Then the cells below the model-output row were filled with data from 1001 recalculations of the simulation model. To speed things up, the computer did not bother to actually display all of these 1001 recalculations as it was doing them, but the values that would have been displayed in the model-output cells B35:C35 have been written into the 1001 rows from row 36 to row 1035. In the first recalculation of the simulation model, the salesperson was unskilled and made 5 sales, and so 0 (for "No" as answer to the question "Skill high?") and 5 are written in cells B36 and C36 (our first data row). In the second recalculation of the model, the salesperson was unskilled and made 9 sales, and so 0 and 9 are written in B37:C37. Continuing down the list, each row in the range B36:C1036 contains the skill level (in the B column) and the number of sales (in the C column) for one recalculation of our simulation model. Notice in Figure 5 that the sales numbers tend to be smaller in the rows where the salesperson does not have high skill. In each row the B

and C numbers are linked by the relationships that are built into our model, but the data in different rows are independent of each other because each row corresponds to another independent recalculation of the whole model from Figure 4.

The data from the 1001 simulations are stored in B36:C1036 as values that do not change when we recalculate the spreadsheet by pressing [F9]. Having these values fixed is important, because it allows us to sort the data and analyze it statistically without having it change every time we calculate another statistic. But above the data range, the model-output cells B35:C35 still contain the formulas that link them to our simulation model, and so these two cells will change when [F9] is pressed.

In the cells A36:A1036, on the left edge of the simulation table, Simtools has entered percentile value numbers that shows, for each row of simulation data, what fraction of the other data rows are above this row. Because we selected a range with 1001 data rows, these percentile numbers increase by 1/1000 per row, increasing from 0 in the first data row (row 36) to 1 in the last data row (row 1036). (Using 1001 data rows gives us nice even 1/1000 increments here because each data row has 1000 "other data rows" above and below it. These percentile values will be used later for making cumulative distribution charts, after sorting the simulation data.)

Now, recall our question about what we can infer about the salesperson's skill level if he gets 9 sales in 20 calls. Paging down the data range, we may find some rows where a skilled salesperson got 9 sales, but most of the cases where 9 sales occurred are cases where the salesperson was unskilled. To get more precise information out of our huge data set, we need to be able to efficiently count the number of times that 9 sales (or any other number of sales) occurred with either skill level.

To analyze the skill levels where in the simulations where a salesperson made exactly 9 sales, let us first enter the number 9 into the cell D34. (Cell D34 will be our comparison cell that we can change if we want to count the number of occurrences of some other number of sales.)

Next, into cell E36, we enter the formula

```
=IF(C36=$D$34,B36,"..")
```

Then we copy the range E36 and paste it to the range E36:E1036. Now the cells in the E column have the value 1 in each data row where a skilled salesperson gets 9 sales, the value 0 in each

data row where an unskilled salesperson gets 9 sales, and the value "." in all other data rows. (Notice the importance of having absolute references to cell \$D\$34 in the above formulas, so that they do not change when they are copied!) With the cells in E36:E1036 acting as counters, we can count the number of times that 9 sales occurred in our simulation data by entering

=COUNT( E36 : E1036 )

into cell E28 (say). In this formula we are using the fact that Excel's COUNT function ignores cells that have non-numerical values (like "." here). We can also use the same E cells to tell us how many times that 9 sales occurred for skilled salesperson, by entering the formula

=SUM( E36 : E1036 )

into cell E27. Here we are using the fact that Excel's SUM function also ignored non-numerical cells.

In my simulation data, the COUNT formula shows nine sales occurred in 68 of our 1001 simulations, while the SUM formula shows that salespeople with high skills were responsible for 13 of these 68 simulations where nine sales occurred. Now, suppose that we have a new salesperson whose skill level is not known to us, but we learn that he also made nine sales in twenty calls. If our uncertainty about the skill and sales totals for this person are like those in our simulation data, then he should look like another draw from the population that gave us these 68 simulated salespeople who made nine sales. Thus, given this sales information, we can estimate that his probability of being skilled is approximately  $13/68 = 0.191$ . This result is computed by the formula =E27/E28 in cell E30 of Figure 6, which returns the same number that we could also get by the formula =AVERAGE( E36 : E1036 ).

This number  $13/68 = 0.191$  in cell E30 may be called our estimated conditional probability of a salesperson having high skill given that he has made nine sales in twenty calls, based on our simulation data. In the notation of probability theory, we often use the symbol " $|$ " to denote the word "given", and mathematicians often use " $\approx$ " to denote the phrase "is approximately equal to". With this notation, the result of our simulation may be summarized by writing  $P(\text{Skill high} | \text{Sales}=9) \approx 13/68$ .

	A	B	C	D	E	F	G
25	FORMULAS FROM RANGE A26:E1036				With Sales=\$D\$34,		
26	B35. =IF(D2=2/3,1,0)				Frequency in Simtable:		
27	C35. =D10				13	Skill hi	
28	E36. =IF(C36=\$D\$34,B36,"..")				68	All	
29	E36 copied to E36:E1036				P(Skill hi Sales=\$D\$34)		
30	E27. =SUM(E36:E1036)				0.191176		
31	E28. =COUNT(E36:E1036)						
32	E30. =E27/E28						
33				Given Sales=			
34	Skill hi?		Sales	9			
35	SimTable	0	9	Skill hi?			
36	0	0	5	..			
37	0.001	0	9	0			
38	0.002	0	4	..			
39	0.003	0	6	..			
40	0.004	1	16	..			
41	0.005	1	12	..			
42	0.006	0	6	..			
43	0.007	0	6	..			
44	0.008	0	3	..			
45	0.009	0	7	..			

**Figure 6. Simulation data and analysis.**

The conditional probability in cell E30 could also have been computed by the formula =AVERAGE(E36:E1036), because Excel's AVERAGE function also ignores non-numerical cells. That is, AVERAGE(E36:E1036) would return the value  $13/68 = 0.191$ , because the range E36:E1036 contains 68 numerical values, of which 13 are ones and the rest are all zeros.

### 3. Analysis using Excel's data-table command

Changing the 9 in cell D34 to other numbers, we can see the outcome frequencies for other sales numbers. But it would be helpful to make a table in which we can see together the results for all sales numbers from 0 to 20. Such a table can be easily made with Excel's data-table command (sometimes called a "what-if table" in other spreadsheet programs).

Here we will learn to use one form of data table called a column-input data table. The structure of a column-input data table is similar to the simulation table that we made in the

previous section, because (as we shall see) Simtools actually uses Excel's column-input data tables to make its simulation tables.

To make our data table, we must begin by putting the output that we want to tabulate into one row, with space underneath to make the table. We want to tabulate the information contained in the frequency numbers in cells E27 and E28 of Figure 6, but there are not enough blank cells underneath these cells to make the data table there. So let us enter the formula  $=E27$  into cell I34, to echo there the number of skilled salespeople who made the given number of sales in cell D34. Next let us enter the formula  $=E28-E27$  into cell J34, to display there the number of unskilled people who made the given number of sales. Then to display the fraction of skilled among the total for the given number of sales, let us enter the formula  $=I34/(I34+J34)$  into cell K34. This range I34:K34 will be the output range at the top of the data table. Underneath, the data table will tabulate the values of these cells as the parameter in cell D34 is adjusted from 9 to other values between 0 and 20.

The other values that we want to substitute into cell D34 must be listed in the column to the left of the output range, in the rows below it. So we must enter the numbers 0 to 20 into the cells from H36 to H56. (To do so quickly, first enter the number 0 into cell H36, and then we can select the range H36:H56 and use the command sequence Edit>Fill>Series, using the Series dialogue-box options: Columns, Linear, Step-Value 1, and Stop-Value 20.)

Now we select the range H34:K55 and use the command sequence Data>Table. When the "Row and Column Input" dialogue box comes up, we leave the "Row Input" entry blank, but we tab to the "Column Input" box and enter cell D34 as the Column Input Cell.

Following this Data>Table command, Excel will compute the data entries into the range I35:K55 as follows. For each row in this range, Excel first takes the value of the cell in column H (the leftmost column in our selected range H34:K55) and enters it into cell D34. Then Excel recalculates the whole spreadsheet. The new values of the output row at the top of the data table I34:K34 are then copied down into the corresponding (I, J, K) cells in this row. When all the cells in I35:K55 have been filled in this way, Excel restores the original contents of the input cell D34 (the value 9). Notice that the data in row 44 of the data table (I44:K44) is identical to the output above in I34:K34, because row 44 is based on the input value of 9 (from H44), which is the actual

current value of cell D34 (as shown previously in Figure 6).

	H	I	J	K	L	M	N	O
32		Frequencies						
33		Skill hi	Skill lo	P(Skill hi   Sales)				
34	Sales:	13	55	0.191176		FORMULAS FROM RANGE H32:K35		
35	0	0	0	#DIV/0!		I34.	=E27	
36	1	0	1	0		J34.	=E28-E27	
37	2	0	9	0		K34.	=I34/(I34+J34)	
38	3	0	25	0		I35:K55. {=TABLE(,D34)}		
39	4	0	49	0				
40	5	0	82	0				
41	6	0	82	0				
42	7	3	102	0.028571				
43	8	5	62	0.074627				
44	9	13	55	0.191176				
45	10	32	29	0.52459				
46	11	55	10	0.846154				
47	12	70	5	0.933333				
48	13	73	2	0.973333				
49	14	82	0	1				
50	15	75	1	0.986842				
51	16	50	0	1				
52	17	23	0	1				
53	18	5	0	1				
54	19	1	0	1				
55	20	0	0	#DIV/0!				

**Figure 7. Data table of results for different numbers of sales.**

If you check the formulas in the cells from I35 to K55, you will find that they all share the special formula  $\{=TABLE(,D34)\}$ . The braces mean that this is an array formula, which Excel has entered into the whole range I35:K55 at once. (Excel will not let you change any one cell in an array; you have to change all or none. To emphasize that these cells together form an array, I have put a border around the data range I35:K55 in Figure 7, using a Format>Cells command.) The TABLE formula tells us that this range contains the data range of a data table that has no row-input cell but has D34 as its column-input cell. But recall that the whole range that we selected before invoking the Data>Table command also included one row above this data range and one column to the left of this data range. The column on the left side of the data table contains the alternative input values that are substituted one at a time into the designated column-

input cell. The row at the top of the data table contains the output values that are tabulated as these alternative substitutions are done.

The conditional probability of any event A given some other event B, denoted by the formula  $P(A|B)$ , is the probability that we would assign to this event A if we learned that the event B occurred. So the ratios in the K column of the data table give us estimates of the conditional probability of a salesperson's being skilled, given the total number of sales that he has made in 20 calls. Notice how these conditional probabilities of being skilled increase from 0 to 1 as the given sales total.

Our simulation table gives us no data in Figure 7 for the extreme cases of 0 or 20 sales. But it is obvious from the entries immediately below cell K35 that finding 0 successes in the 20 trials should make us almost sure that the salesperson is unskilled. Similarly, the entries just above cell K55 indicate clearly that 20 successes out of 20 trials should make us almost sure that the salesperson is skilled.

[Now I can tell you how you could make a simulation table if you did not have the Simtools add-in for Excel. Recall that we prepared to make our simulation table (as shown in Figures 5 and 6) by putting our model output in cells B35 and C35, and then we selected a range that had these cells plus the cell to their left (cell A35) as its top row, and that also included 1000 more rows below. With this range A35:C1036 selected, we used the command sequence Tools>Simtools>SimulationTable. Now (after saving the file), let us try it again; but this time, after selecting the range A35:C1036, let us use instead the command sequence Data>Table. When the dialogue box asks us to specify input cells, we should leave the row-input box blank, and we should specify a column-input cell that has no effect on any of our calculations (cell A34 or cell A35 will do). The result will be that Excel then recalculates the model 1001 times and stores the results in the rows of B36:C1036, just as before (recall Figures 5 and 6). The column-input substitutions affect nothing, but the recalculation of the RANDs gives us different results in each row. Unfortunately, Excel data tables are alive, in the sense that they will be recalculated every time we recalculate the spreadsheet. To do statistical analysis, we really do not want our simulation data to keep changing! To fix this, we should select the data range B36:C1036, copy it to the clipboard (by Edit>Copy), and then with B36:C1036 still selected we should use the

command sequence

    Edit>PasteSpecial>Values

The result is that the TABLE formulas in the data range are replaced by the values that were displayed, and these numerical values now will not change when [F9] is pressed. The Simtools simulation-table procedure, as written in Excel's VBA macro language, actually tells Excel to do just such a data-table command followed by a copy and paste-special-values command.]

#### 4. Conditional independence

Consider again the spreadsheet model in Figure 4. In this simulation model, the results of the twenty sales calls in the cells B2:B21 are not independent, because they all depend on the random skill level in cell D2. But notice these results also depend on some independent random factors in cells A2:A21, which represent the different customers idiosyncratic feelings about our salesperson's product. In this case, we may say that the results of the 20 sales calls are conditionally independent of each other when the salesperson's skill level is given.

In general, when we say that some random variables, say **X** and **Y**, are conditionally independent given some other random variables, say **Z**, we mean that once you learned the value of **Y**, getting further information about **X** would not affect your beliefs about **Y**, and getting further information about **Y** would not affect your beliefs about **X**. In a spreadsheet model, such conditional independence holds among random cells **X** and **Y** if the random cells **X** and **Y** are not both influenced by any random cells other than **Z**.

Conditional independence is an important and subtle idea. Because the results of the 20 sales calls are not independent in our model (Figure 4), learning that the results of the first 19 calls could cause us to revise our beliefs about the probability of a successful sale resulting from the 20th call. But because the sales calls are conditionally independent given the skill level in this model, if we knew that the salesperson had a high skill level then we would think that his probability of making a sale in the 20th call was 2/3, even if he had not made a sale in any of the first 19 calls.

These concepts of conditional probability and conditional independence will be very important for describing what we do in our spreadsheet simulation models. With this

terminology, the analysis of our model in Section 2 can be summarized as follows:

The salesperson may be either skilled or unskilled, each with probability  $1/2$ . Each of 20 sales calls may result in either a sale or no sale. The results of 20 sales calls are conditionally independent of each other given the salesperson's level of skill. In each call, the conditional probability of a sale would be  $2/3$  given that the salesperson is skilled, but the conditional probability of a sale would be  $1/3$  given that the salesperson is unskilled. Given this situation, we analyzed data from 1001 simulations of the model to estimate that the conditional probability that the salesperson is skilled, given 9 sales in the 20 calls, would be approximately 0.2.

##### 5. A continuous random skill variable from a Triangular distribution

In this course, we are learning how to make quantitative models of managerial situations that involve uncertainty. Even when our models seem complicated, they will always be simplifications which omit or distort much of the real situation that we are trying to study. Of course, it is not possible to think about anything in the world without simplifying it. As powerful as our brains and computers may be, the complexity of the real world is greater than their limited capacity. So perhaps we should not worry too much about simplification. We can look for useful insights from our analysis of a model like that in Figure 4 above, even while recognizing that it is an extreme simplification of the real situation which exists when a new salesperson begins to build a reputation for skill (or lack thereof) in his first sales calls.

But of course there is always a danger that we may have oversimplified and omitted from our model some important details of the real situation that would significantly change our results. So in applied analytical work, we should always be ready to look at more than one variation on our model, where each new variation is an attempt to add another fact of the real world into a model that omitted it.

With this preface, let me say that, although I think that a model like the one in Figure 4 above can give useful insights into a real situation of evaluating salespeople who have limited track records, nevertheless I can imagine that a real sales manager might be disturbed by some of the extreme simplifications that we used to make the model tractable. If I were a consultant to

this manager, my response would be first to get a list of areas where the manager finds shortcomings in my model, and then to build one or more new models that make include her suggestions for more realism in these areas. That is, although I can never make one perfect model, but I should always be prepared to make a sequence of models where each new model addresses specific concerns that have been expressed about my previous models. To do so, I need to have a versatile toolkit for making analytical models. The ultimate goal of our study here is of course to develop such a toolkit.

So now let us suppose that a sales manager looks at our model in Figure 4 and describes the assumption that there are only two possible levels of skill as an absurd oversimplification. If his willing to accept our concept of a "skill level" that represents a potential long-run rate of success in sales calls, he may tell us that a salesperson's "skill level" in this sense could be any number between 0 and 1, not just 1/3 or 2/3. We might then ask the manager to describe more fully her beliefs about new salespeople before they make their first twenty sales calls. Suppose that, in response, the manager repeats that a salesperson's potential long run success rate could be anywhere between 0% and 100%, but adds that success rates close to 50% are probably the most common. So to take account of these beliefs, we should represent the skill level by ar random variable random variable that can take any value between 0 and 1, and that is more likely to be near 0.5 than any other number.

A unknown quantity that can take any possible value in an interval of numbers is called a continuous random variable. To generate such continuous random variables, we generally use one of several famous mathematical probability formulas. Among the commonly used formulas, one of the simplest formulas that can be used to model beliefs like those described above is the formula for a Triangular random variable. We can generate such Triangular random variables in our spreadsheets by a Simtools function called TRIANINV. To learn about this statistical function, use the command sequence Insert>Function to launch the PasteFunction dialogue box, and then search for TRIANINV among the Statistical functions. You will find that TRIANINV takes four parameters. After the first parameter, the latter three parameters are called "lowerbound," "mostlikely," and "upperbound." So the second, third, and fourth parameters should be 0, 0.5, and 1 when we want to make a random variable that could be any number from

0 to 1 but is most likely to be near 0.5. The first parameter, which is mysteriously called "probability," should be a RAND() function, to make the result a random variable. Thus, the formula

$$= \text{TRIANGINV}(\text{RAND}(), 0, 0.5, 1)$$

returns a Triangular random variable that could be any number from 0 to 1 but is more likely to be near 0.5 than any other number in this interval.

What does a Triangular random variable really mean? I could tell you that its probability density has a simple triangular shape, positive over the interval from 0 to 1, with a peak at 0.5. But if that does not mean much to you, you should just enter this formula into cell D2 of our spreadsheet, press [F9] many times, and watch how this value jumps around. It can take any value between 0 and 1, but the first decimal place is more likely to be 4 or 5 than any other digit. So our uncertainty about the next value of this formula should be a good model for simulating the manager's beliefs about the new salesperson.

Figure 8 shows a revised version of our old model from Figure 4, in which the salesperson's simulated skill level in cell D2 has been changed to a triangular random variable, generated by the formula  $=\text{TRIANGINV}(\text{RAND}(), H2, H3, H4)$ , where the value of cell H2 is the lower bound 0, the value of H3 is the most likely point 0.5, and the value of cell H4 is the upper bound 1 for the random variable.

	A	B	C	D	E	F	G	H	I
1	(rands)	Sales:		Salesperson's level of skill (Triangular)					
2	0.30521	1		0.41471		lower bound		0	
3	0.681339	0				most likely		0.5	
4	0.637888	0				upper bound		1	
5	0.909307	0							
6	0.511003	0							
7	0.772921	0							
8	0.069371	1							
9	0.504553	0		Total Sales					
10	0.424588	0		9					
11	0.08702	1							
12	0.836577	0							
13	0.231236	1							
14	0.179568	1							
15	0.746769	0							
16	0.226163	1			FORMULAS FROM RANGE A1:D21				
17	0.35638	1			A2.	=RAND()			
18	0.218511	1			B2.	=IF(A2<=\$D\$2,1,0)			
19	0.211323	1			A2:B2 copied to A3:A21				
20	0.612002	0			D2.	=TRIANINV(RAND(),H2,H3,H4)			
21	0.681832	0			D10.	=SUM(B2:B21)			

**Figure 8. Sales model with skill level from a Triangular distribution.**

Now we need to analyze the model and find how our beliefs about the salesperson should change after we observe the number of successful sales he gets in 20 calls. In our simulation data, we need to tabulate the skill level (from cell D2) and the number of successful sales (from cell D10) for each simulation. To echo the values of these random variables in one row, to be the top of our simulation table, let us enter the formula =D2 in cell B35 and the formula =D10 in cell C35. Then let us select the range A35:C1036 and enter the command sequence Tools>Simtools>SimulationTable. When Excel finishes calculating, the skill and sales numbers from 1001 simulations are contained in the rows of the data range B36:C1036.

Now we want to find, what are the typical skill levels of people who get 9 sales out of 20 such calls? One way to gather this information is to select the range B36:C1036 and use the command sequence Data>Sort to sort this data range by the numbers of sales in the C column. (In the Sort dialogue box, it may be helpful to indicate that the selected list has no header row.) Then the data from all simulations where 9 sales occurred would be gathered together in one

block of about a seventy rows (probably beginning near row 400), and we could simply look down to see what skill levels can be found there. But I want to show other ways to do this conditional data analysis, using cell formulas instead of sorting or filtering the spreadsheet, because using cell formulas will ultimately give you more analytical power and versatility.

Our basic technique is to create one or more columns like E36:E1036 in Figure 6, where numerical data is taken only from the rows where we find the number of sales that we want to study. So we begin as before by entering the number 9 into cell D34, to represent the number of sales that we want to study. Then we can enter the formula

$$=IF(C36=\$D\$34,B36,"..")$$
 into cell E36,

and then we can copy E36 to the range E36:E1036. Now the E column is selecting for us all the skills of simulated salespeople who sold to nine customers.

The number of simulations in which the number of successful sales was exactly nine is computed in cell I27 of Figure 9 by the formula

$$=COUNT(E36:E1036)$$

(Recall that the COUNT function ignores nonnumerical cells.)

Each of these 80 numerical cells in E36:E1036 contains a different skill level. To understand something about the distribution of skills for people who get nine successful sales, we may want to count how many of these cells contain skill levels less than (say) 0.5 or any other value  $x$ . So let us enter the number 0.5 into cell F31. Then the number of numerical cells in E36:E1036 that have values less than or equal to the value of cell F31 can be calculated by the Excel formula

$$=FREQUENCY(E36:E1036,F31)$$

Cell I31 in Figure 9 contains the formula

$$=FREQUENCY(E36:E1036,F31)/I27$$

which returns the fraction (0.65) of cases where the skill level was less than 0.5 among all those cases where the number of sales was 9. So based on this simulation data, we can estimate that 0.65 is the conditional probability of having skill less than 0.5 given that there were nine successful sales out of twenty calls.

More generally, if we want to estimate the conditional probability that the skill level is less

than any number x, given that the number of sales was any number m, then we only have to enter the number x into cell F31 and the number m into cell D34 in Figure 9. The desired conditional probability (as estimated from our simulation data) will be given by the FREQUENCY(E36:E1036,F31)/COUNT(E36:E1036) value shown in cell I31.

	A	B	C	D	E	F	G	H	I	J
22	FORMULAS FROM RANGE A30:J1036									
23	B35. =D2		C35. =D10							
24	E36. =IF(C36=\$D\$34, B36, "..")									
25	E36 copied to E36:E1036						Simulation sample size			
26	I27. =COUNT(E36:E1036)						with Sales=D34		Total	
27	J27. =COUNT(B36:B1036)							80	1001	
28	I31. =FREQUENCY(E36:E1036,F31)/I27									
29	J31. =FREQUENCY(B36:B1036,F31)/J27						Cumulative probabilities			
30	I37. =PERCENTILE(\$E\$36:\$E\$1036,H37)				x	P(Sk<x Sales=D34)			P(Sk<x)	
31	J37. =PERCENTILE(\$B\$36:\$B\$1036,H37)				0.5			0.65	0.4955	
32	I37:J37 copied to I37:J57									
33				Sales=						
34		Skill	Sales	9						
35	SimTable	0.6639	14		Skill				Skill level	
36	0	0.61	15		..	CumulativePr		Sales	Prior	
37	0.001	0.4375	7		..		0	0.2285	0.01087	
38	0.002	0.6598	12		..		0.05	0.3212	0.15678	
39	0.003	0.3178	10		..		0.1	0.3498	0.21831	
40	0.004	0.3072	8		..		0.15	0.3623	0.26667	
41	0.005	0.8484	14		..		0.2	0.3759	0.30838	
42	0.006	0.4794	9		0.4794		0.25	0.3937	0.34537	
43	0.007	0.4642	9		0.4642		0.3	0.4127	0.38285	
44	0.008	0.5559	8		..		0.35	0.4361	0.42112	
45	0.009	0.785	15		..		0.4	0.4475	0.44944	
46	0.01	0.475	6		..		0.45	0.4607	0.47838	
47	0.011	0.6562	13		..		0.5	0.4769	0.50482	
48	0.012	0.1883	7		..		0.55	0.4841	0.53004	
49	0.013	0.4051	11		..		0.6	0.4927	0.55128	
50	0.014	0.7897	16		..		0.65	0.5007	0.57886	
51	0.015	0.474	12		..		0.7	0.5218	0.60848	
52	0.016	0.2511	3		..		0.75	0.5388	0.64954	
53	0.017	0.7013	15		..		0.8	0.5567	0.68238	
54	0.018	0.2972	4		..		0.85	0.5752	0.73366	
55	0.019	0.491	10		..		0.9	0.5868	0.76867	
56	0.02	0.6656	15		..		0.95	0.6196	0.84567	
57	0.021	0.6824	14		..		1	0.7396	0.97366	

Figure 9. Simulation data and analysis for continuous-skill model.

Now let's turn the question around. Instead specifying a skill level  $x$  and asking what is the probability of skill being lower than  $x$ , we could instead specify a probability  $p$  and ask, what skill level has this much probability below it? The analysis in H36:I57 of Figure 9 addresses this question.

The numbers 0, 0.05, 0.10,... 0.95, 1 are listed in cells H36:H57. (I entered these values by first entering 0 into H36, then selecting H36:H57, and then using the command sequence Edit>Fill>Series, with Columns, Linear, Step=0.05, and Stop=1 selected in the subsequent Series dialogue box.) Then the formula

$$=PERCENTILE(\$E\$36:\$E\$1036, H37)$$

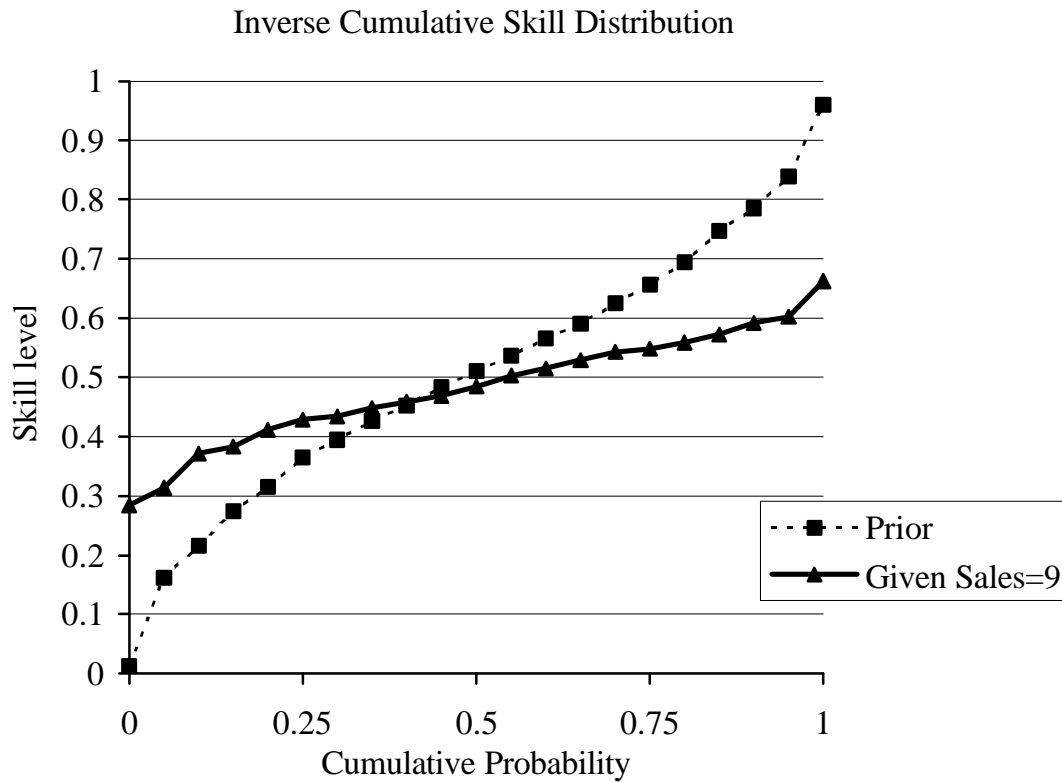
is entered into cell I37, and I37 is copied to I37:I57.

To see what these percentile numbers mean, let us look first at cell I42, for example. The formula here is  $=PERCENTILE(\$E\$36:\$E\$1036, H42)$ , and the value of cell H42 is 0.25. So this formula answers the question: For what number  $x$  will we find that 25% of the numbers in E36:E1036 are less than  $x$ ? According to cell I42, the answer is  $p=0.3937$ . But now remember that E36:E1036 contains skill numbers only for the simulations where nine sales occurred (as long as cell D34 has the value 9). So the value 0.3937 in cell I42 tells us that, among those simulations where nine sales occurred, about 25% had skill levels less than 0.3937.

Similarly, the value of  $PERCENTILE(E36:E1036, 0.5)$  in cell I47 tells us that, among those simulations where nine sales occurred, about 50% had skill levels less than 0.4769. The value of  $PERCENTILE(E36:E1036, 0.75)$  in cell I52 tells us that, among those simulations where nine sales occurred, about 75% had skill levels less than 0.5388. The value of  $PERCENTILE(E36:E1036, 0)$  in cell I37 is the smallest numerical value in the range E36:E1036, and the value of  $PERCENTILE(E36:E1036, 1)$  in cell I57 is the highest numerical value in the range E36:E1036. So cells I37 and I57 tell us that, in those simulations where nine sales occurred, the lowest skill level was 0.2285 and the highest skill level was 0.7396.

For comparison, cells J37:J57 in Figure 9 show the corresponding percentile values in the whole simulation sample that was generated by the Triangular distribution. Cell J37 contains the formula  $=PERCENTILE(\$B\$36:\$B\$1036, H42)$ , and this cell has been copied to J37:J57. The list of these percentile values from the whole unfiltered sample may be called the prior

distribution for the unknown skill level, because they represent what we believed before we learned how many sales the person got in the 20 calls.



**Figure 10. Prior and posterior skill distributions, estimated from simulation data.**

Figure 10 shows the results of all the percentile formulas in I37:J57. You can make such a chart from the spreadsheet in Figure 9 by selecting the range H37:J57 and then using the command sequence Insert>Chart>XY-Scatter. A chart like Figure 10 is sometimes called an inverse cumulative probability chart. In mathematical writing, the probability of an unknown quantity being less than x is often called the cumulative probability of x. In the past, when mathematicians made charts showing the cumulative probabilities of various values, they usually preferred to put the cumulative probabilities on the vertical axis, and so they would use the term "inverse" for our Figure 10, where the cumulative probabilities are on the horizontal axis. But in this course we will always find it more convenient and natural to put the cumulative probabilities

on the horizontal axis, and so we may drop this term "inverse" when we talk about making such charts. (Mathematicians similarly use the phrase inverse cumulative probability function to refer to a function that, for any number  $p$  between 0 and 1, returns the value  $x$  such that some random variable has probability  $p$  of being less than  $x$ . TRIANINV is actually such a function, and this is the reason for the ending "INV" in its name.)

You could also make a histogram to represent the frequency of various skill levels given that 9 sales occurred. But with practice you should learn to see such probability information better in (inverse) cumulative charts like Figure 10. The lowest likely values of the unknown quantity under consideration (here the skill levels of people who have made 9 sales) are represented by the height of the inverse-cumulative curve at its left end, at cumulative probabilities near 0. The highest possible value of this unknown quantity are represented by the height of the curve at its right side, at cumulative probabilities near 1. The height of the curve in the middle, at cumulative probability 0.5, is the median which the unknown quantity is equally likely to be above or below. If you look at the height of the curve at cumulative probability 0.25 and 0.75, then you get a pair of values that the unknown quantity may be between with probability 1/2. For an interval that has a 90% probability of including the unknown quantity, let the lower end be the height of the curve at cumulative probability 0.05, and let the upper end be the height of the curve at cumulative probability 0.95. Figure 10 shows that, after 9 sales have been observed, this 90% probability interval for the unknown skill level goes from about 0.32 to 0.63. In the prior, before the number of sales was observed, the Triangular distribution that has been assumed here would have 90% probability in the interval from about 0.16 to 0.84 instead. An overall comparison of these two curves in Figure 10 shows that the observation of 9 sales changes our beliefs about his skill level by increasing in the perceived worst-case low end somewhat while decreasing the best-case high end substantially more. After all, his performance could have been somewhat worse, but it also could have been much better.

## 6. Summary

This chapter focused on an example in which we want to learn about some unknown quantity (the salesperson's skill level) by observing other events (the successes and failures in

various sales calls) that are influenced by this unknown quantity. We analyzed this problem using spreadsheet simulation models in which one random cell simulates the unknown quantity and other cells simulate the observations that depend on the unknown quantity.

In the context of this problem, we introduced some basic ideas of probability and some basic techniques of spreadsheet modeling. Probability ideas introduced here include: prior and conditional probabilities  $P(A|B)$ , independence, conditional independence, uniform and triangular probability distributions, cumulative probabilities, and (inverse) cumulative probability charts. Excel functions introduced here include: RAND, IF, COUNT, SUM, AVERAGE, PERCENTILE, and FREQUENCY. We also introduced the Simtools function TRIANINV. We saw how to get information about these and other technical functions in Excel, by the using of the insert-function dialogue box. Other basic spreadsheet techniques used in this chapter include: absolute (\$) and relative references in formulas, simulation tables, column-input data tables, filled series, the paste-special-values command, and the basic XY-Scatter chart.

To compute conditional probabilities from large tables of simulation data, we introduced a formula-filtering technique in which a column is filled with IF formulas that extract information from a simulation table, returning a nonnumerical value ("..") in data rows that do not match our criterion. The information extracted in such columns was summarized using statistical functions like COUNT, SUM, AVERAGE, FREQUENCY, and PERCENTILE, which are designed to ignore nonnumerical entries.

## Problems for Introduction to Probability and Simulation in Spreadsheets

1. The Connecticut Electronics company produces sophisticated electronic modules in production runs of several thousand at a time.. It has been found that the fraction of defective modules in can be very different in different production runs. These differences are caused by micro-irregularities that sometimes occur in the electrical current. For a simple first model, we may assume first that there are just two possible values of the defective rate.

In about 70% of the production runs, the electric current is regular, in which case every module that is produced has an independent 10% chance of being defective.

In the other 30% of production runs, when current is irregular, every module that is produced has an independent 40% chance of being defective.

Testing these modules is quite expensive, so it is valuable to make inferences about the overall rate of defective output based on a small sample of tested modules from each production run.

(a) Make a spreadsheet model to study the conditional probability of irregular current given the results of testing 10 modules from a production run. Make a simulation table with data from at least 1000 simulations of your model (where each simulation includes the results of testing 10 modules), and use this table to answer the following questions.

(b) Before testing any modules, what is the probability of finding exactly 2 defective modules among the 10 tested?

(c) What is the conditional probability that the current is regular given that 2 defective modules are found among the 10 tested?

(d) Make a table showing the estimates of  $P(\text{irregular current} | k \text{ defectives among 10 tested})$ , for  $k = 0, 1, 2, \dots$ . (You may have trouble with some  $k$  greater than 7, but the answer in those cases should be clear.)

2. Reconsider the Connecticut Electronics problem, but let us now drop the assumption that each run's defective rate must be one of only two possible values. Managers have observed that, due to differences in the electrical current, the defective rate in any production run may be any number between a lower bound of 0% and an upper bound of 50%. They have also observed that the defective rates are more likely to be near 10% than any other single value in this range. So suppose that the defective rate on any production run is drawn from a Triangular distribution with these parameters. We want to quantify what we can infer about the defective rate in the most recent production run based on the testing of ten modules from this production run.

(a) Before we test any modules, what is the probability of a defective rate less than 0.25 in this production run? What is the median defective rate for such a run?

(b) Based on data from at least 1000 simulations, what is the probability that we will find exactly 2 defective modules when we test ten modules from this production run?

(c) If we find exactly 2 defective modules when we test ten modules from this production run then, given these results, what is the median value for this production run? Show the (inverse) cumulative distribution for this run's defective rate, given 2 defective among ten tested.

(d) To your chart for part (c), add two more curves to show the cumulative distributions for this run's defective rate if the number of defective modules among ten tested is 1, and if it is 3.

\*(e) Make a table listing, for any integer  $k$  from 0 to 10, the conditional expected value of this run's overall defective rate, given that  $k$  defective modules are found in a sample of ten modules from the run.