PART III

Gathering Data

Chapter 11
Understanding Randomness

Chapter 12
Sample Surveys

Chapter 13
Experiments and Observational Studies

CHAPTER 11

Understanding Randomness

"The most decisive conceptual event of twentieth century physics has been the discovery that the world is not deterministic. . . . A space was cleared for chance."

— Ian Hocking, *The Taming of Chance*

We all know what it means for something to be random. Or do we? Many children's games rely on chance outcomes. Rolling dice, spinning spinners, and shuffling cards all select at random. Adult games use randomness as well, from card games to lotteries to Bingo. What's the most important aspect of the randomness in these games? It must be fair.

What is it about random selection that makes it seem fair? It's really two things. First, nobody can guess the outcome before it happens. Second, when we want things to be fair, usually some underlying set of outcomes will be equally likely (although in many games, some combinations of outcomes are more likely than others).

Randomness is not always what we might think of as "at random." Random outcomes have a lot of structure, especially when viewed in the long run. You can't predict how a fair coin will land on any single toss, but you're pretty confident that if you flipped it thousands of times you'd see about 50% heads. As we will see, randomness is an essential tool of Statistics. Statisticians don't think of randomness as the annoying tendency of things to be unpredictable or haphazard. Statisticians use randomness as a tool. In fact, without deliberately applying randomness, we couldn't do most of Statistics, and this book would stop right about here.[1]

But truly random values are surprisingly hard to get. Just to see how fair humans are at selecting, pick a number at random from the top of the next page. Go ahead. Turn the page, look at the numbers quickly, and pick a number at random.

Ready?

Go.

[1] Don't get your hopes up.

1 2 3 4

It's Not Easy Being Random

Did you pick 3? If so, you've got company. Almost 75% of all people pick the number 3. About 20% pick either 2 or 4. If you picked 1, well, consider yourself a little different. Only about 5% choose 1. Psychologists have proposed reasons for this phenomenon, but for us, it simply serves as a lesson that we've got to find a better way to choose things at random.

"The generation of random numbers is too important to be left to chance."

—Robert R. Coveyou, Oak Ridge National Laboratory

So how should we generate **random numbers?** It's surprisingly difficult to get random values even when they're equally likely. Computers have become a popular way to generate random numbers. Even though they often do much better than humans, computers can't generate truly random numbers either. Computers follow programs. Start a computer from the same place, and it will always follow exactly the same path. So numbers generated by a computer program are not truly random. Technically, "random" numbers generated this way are *pseudorandom* numbers. Pseudorandom values are generated in a fixed sequence, and because computers can represent only a finite number of distinct values, the sequence of pseudorandom numbers must eventually repeat itself. Fortunately, pseudorandom values are good enough for most purposes because they are virtually indistinguishable from truly random numbers.

***Activity:* Random Behavior.** *ActivStats*' Random Experiment Tool lets you experiment with truly random outcomes. We'll use it a lot in the coming chapters.

TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
scottadams@aol.com
www.dilbert.com
NINE NINE NINE NINE NINE NINE
10/25/01 © 2001 United Feature Syndicate, Inc.
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RANDOMNESS: YOU CAN NEVER BE SURE.

***Activity:* Truly Random Values on the Internet.** This activity will take you to an Internet site (www.random.org) that generates all the truly random numbers you could want.

There *are* ways to generate random numbers so that they are both equally likely and truly random. In the past, entire books of carefully generated random numbers were published. The books never made the best-seller lists and probably didn't make for great reading, but they were quite valuable to those who needed truly random values.[2] Today, we have a choice. We can use these books or find genuinely random digits from several Internet sites. The sites use methods like timing the decay of a radioactive element or even the random changes of lava

[2] You'll find a table of random digits of this kind in the back of this book.

lamps to generate truly random digits.[3] In either case, a string of random digits might look like this:

```
2217726304387410092537086270581997622725849795907032825001108963
3217535822643800292254644943760642389043766557204107354186024508
8906427308645681412198226653885873285801699027843110380420067664
8740522639824530519902027044464984322000946238678577902639002954
8887003319933147508331265192321413908608674496383528968974910533
6944182713168919406022181281304751019321546303870481407676636740
6070204916508913632855351361361043794293428486909462881431793360
7706356513310563210508993624272872250535395513645991015328128202
```

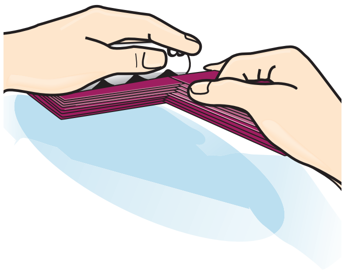
An ordinary deck of playing cards, like the ones used in bridge and many other card games, consists of 52 cards. There are numbered cards (2 through 10), and face cards (Jack, Queen, King, Ace) whose value depends on the game you are playing. Each card is also marked by one of four suits (clubs, diamonds, hearts, or spades) whose significance is also game-specific.

You probably have more interesting things to download than a few million random digits, but we'll discuss ways to use such random digits to apply randomness to real situations soon. The best ways we know to generate data that give a fair and accurate picture of the world rely on randomness, and the ways in which we draw conclusions from those data depend on the randomness, too.

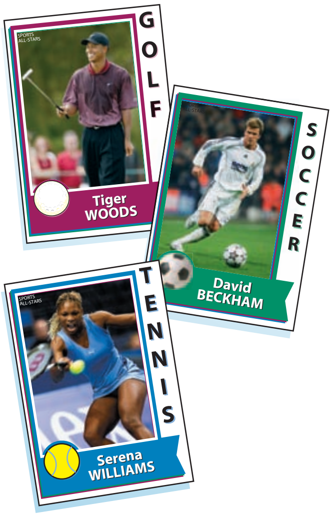
Aren't you done shuffling yet? Even something as common as card shuffling may not be as random as you might think. If you shuffle cards by the usual method in which you split the deck in half and try to let cards fall roughly alternately from each half, you're doing a "riffle shuffle."

How many times should you shuffle cards to make the deck random? A surprising fact was discovered by statisticians Persi Diaconis, Ronald Graham, and W. M. Kantor. It takes seven riffle shuffles. Fewer than seven leaves order in the deck, but after that, more shuffling does little good. Most people, though, don't shuffle that many times.

When computers were first used to generate hands in bridge tournaments, some professional bridge players complained that the computer was making too many "weird" hands—hands with 10 cards of one suit, for example. Suddenly these hands were appearing more often than players were used to when cards were shuffled by hand. The players assumed that the computer was doing something wrong. But it turns out that it's humans who hadn't been shuffling enough to make the decks really random and have those "weird" hands appear as often as they should.



Practical Randomness



Suppose a cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal in the hope of boosting sales. The manufacturer announces that 20% of the boxes contain a picture of Tiger Woods, 30% a picture of David Beckham, and the rest a picture of Serena Williams. You want all three pictures. How many boxes of cereal do you expect to have to buy in order to get the complete set?

How can we answer questions like this? Well, one way is to buy hundreds of boxes of cereal to see what might happen. But let's not. Instead, we'll consider using a random model. Why random? When we pick a box of cereal off the shelf, we don't know what picture is inside. We'll assume that the pictures are randomly placed in the boxes and that the boxes are distributed randomly to stores around the country. Why a model? Because we won't actually buy the cereal boxes. We can't afford all those boxes and we don't want to waste food. So we need an imitation of the real process that we can manipulate and control. In short, we're going to **simulate** reality.

[3] For example, www.random.org or www.randomnumbers.info.

A Simulation

Modern physics has shown that randomness is not just a mathematical game; it is fundamentally the way the universe works.

Regardless of improvements in data collection or in computer power, the best we can ever do, according to quantum mechanics . . . is predict the probability that an electron, or a proton, or a neutron, or any other of nature's constituents, will be found here or there. Probability reigns supreme in the microcosmos.
—Brian Greene, *The Fabric of the Cosmos: Space, Time, and the Texture of Reality* (p. 91)

The question we've asked is how many boxes do you expect to buy to get a complete card collection. But we can't answer our question by completing a card collection just once. We want to understand the *typical* number of boxes to open, how that number varies, and, often, the shape of the distribution. So we'll have to do this over and over. We call each time we obtain a simulated answer to our question a **trial.**

For the sports cards, a trial's outcome is the number of boxes. We'll need at least 3 boxes to get one of each card, but with really bad luck, you could empty the shelves of several supermarkets before finding the card you need to get all 3. So, the possible outcomes of a trial are 3, 4, 5, or lots more. But we can't simply pick one of those numbers at random, because they're not equally likely. We'd be surprised if we only needed 3 boxes to get all the cards, but we'd probably be even more surprised to find that it took exactly 7,359 boxes. In fact, the reason we're doing the simulation is that it's hard to guess how many boxes we'd expect to open.

Building a Simulation

We know how to find equally likely random digits. How can we get from there to simulating the trial outcomes? We know the relative frequencies of the cards: 20% Tiger, 30% Beckham, and 50% Serena. So, we can interpret the digits 0 and 1 as finding Tiger; 2, 3, and 4 as finding Beckham; and 5 through 9 as finding Serena to simulate opening one box. Opening one box is the basic building block, called a **component** of our simulation. But the component's outcome isn't the result we want. We need to observe a sequence of components until our card collection is complete. The *trial's* outcome is called the **response variable**; for this simulation that's the *number* of components (boxes) in the sequence.

Let's look at the steps for making a simulation:

Specify how to model a component outcome using equally likely random digits:

1. **Identify the component to be repeated.** In this case, our component is the opening of a box of cereal.
2. **Explain how you will model the component's outcome.** The digits from 0 to 9 are equally likely to occur. Because 20% of the boxes contain Tiger's picture, we'll use 2 of the 10 digits to represent that outcome. Three of the 10 digits can model the 30% of boxes with David Beckham cards, and the remaining 5 digits can represent the 50% of boxes with Serena. One possible assignment of the digits, then, is

   0, 1 Tiger   2, 3, 4 Beckham   5, 6, 7, 8, 9 Serena.

Specify how to simulate trials:

3. **Explain how you will combine the components to model a trial.** We pretend to open boxes (repeat components) until our collection is complete. We do this by looking at each random digit and indicating what picture it represents. We continue until we've found all three.
4. **State clearly what the response variable is.** What are we interested in? We want to find out the number of boxes it might take to get all three pictures.

Put it all together to run the simulation:

5. **Run several trials.** For example, consider the third line of random digits shown earlier (p. 257):

   8906427308645681412198226653885873285801699027843110380420067664.

   Let's see what happened.

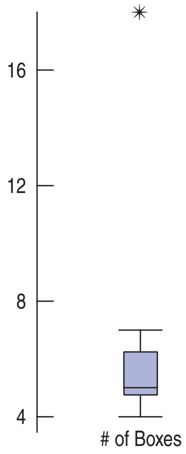
The first random digit, 8, means you get Serena's picture. So the first component's outcome is Serena. The second digit, 9, means Serena's picture is also in the next box. Continuing to interpret the random digits, we get Tiger's picture (0) in the third, Serena's (6) again in the fourth, and finally Beckham (4) on the fifth box. Since we've now found all three pictures, we've finished one trial of our simulation. This trial's outcome is 5 boxes.

Now we keep going, running more trials by looking at the rest of our line of random digits:

89064 2730 8645681 41219 822665388587328580 169902 78431 1038 042006 7664.

It's best to create a chart to keep track of what happens:

Trial Number	Component Outcomes	Trial Outcomes: y = Number of boxes
1	89064 = **Serena**, Serena, **Tiger**, Serena, **Beckham**	5
2	2730 = **Beckham**, **Serena**, Beckham, **Tiger**	4
3	8645681 = **Serena**, Serena, **Beckham**, . . . , **Tiger**	7
4	41219 = **Beckham**, **Tiger**, Beckham, Tiger, **Serena**	5
5	822665388587328580 = **Serena**, **Beckham**, . . . , **Tiger**	18
6	169902 = **Tiger**, **Serena**, Serena, Serena, Tiger, **Beckham**	6
7	78431 = **Serena**, Serena, **Beckham**, Beckham, **Tiger**	5
8	1038 = **Tiger**, Tiger, **Beckham**, **Serena**	4
9	042006 = **Tiger**, **Beckham**, Beckham, Tiger, Tiger, **Serena**	6
10	7664 . . . = **Serena**, Serena, Serena, **Beckham** . . .	?



A S *Activity:* **Bigger Samples Are Better.** The random simulation tool can generate lots of outcomes with a single click, so you can see more of the long run with less effort.

Analyze the response variable:

6. **Collect and summarize the results of all the trials.** You know how to summarize and display a response variable. You'll certainly want to report the shape, center, and spread, and depending on the question asked, you may want to include more.
7. **State your conclusion,** as always, in the context of the question you wanted to answer. Based on this simulation, we estimate that customers hoping to complete their card collection will need to open a median of 5 boxes, but it could take a lot more.

If you fear that these may not be accurate estimates because we ran only nine trials, you are absolutely correct. The more trials the better, and nine is woefully inadequate. Twenty trials is probably a reasonable minimum if you are doing this by hand. Even better, use a computer and run a few hundred trials.

FOR EXAMPLE Simulating a dice game

The game of 21 can be played with an ordinary 6-sided die. Competitors each roll the die repeatedly, trying to get the highest total less than or equal to 21. If your total exceeds 21, you lose.

Suppose your opponent has rolled an 18. Your task is to try to beat him by getting more than 18 points without going over 21. How many rolls do you expect to make, and what are your chances of winning?

Question: How will you simulate the components?

A component is one roll of the die. I'll simulate each roll by looking at a random digit from a table or an Internet site. The digits 1 through 6 will represent the results on the die; I'll ignore digits 7–9 and 0.

(*continued*)

For Example (*continued*)

Question: How will you combine components to model a trial? What's the response variable?

I'll add components until my total is greater than 18, counting the number of rolls. If my total is greater than 21, it is a loss; if not, it is a win. There are two response variables. I'll count the number of times I roll the die, and I'll keep track of whether I win or lose.

Question: How would you use these random digits to run trials? Show your method clearly for two trials.

91129 58757 69274 92380 82464 33089

I've marked the discarded digits in color.

Trial #1:	9	1	1	2	9	5	8	7	5	7	6	
Total:		1	2	4		9			14		20	Outcomes: 6 rolls, won

Trial #2:	9	2	7	4	9	2	3	8	0	8	2	4	6	
Total:		2		6		8	11				13	17	23	Outcomes: 7 rolls, lost

Question: Suppose you run 30 trials, getting the outcomes tallied here. What is your conclusion?

Based on my simulation, when competing against an opponent who has a score of 18, I expect my turn to usually last 5 or 6 rolls, and I should win about 70% of the time.

Number of rolls	
4	|||
5	𝍸 𝍸
6	𝍸 𝍸 |
7	𝍸
8	|

Result	
Won	𝍸 𝍸 𝍸 𝍸 |
Lost	𝍸 ||||

JUST CHECKING

The baseball World Series consists of up to seven games. The first team to win four games wins the series. The first two are played at one team's home ballpark, the next three at the other team's park, and the final two (if needed) are played back at the first park. Records over the past century show that there is a home field advantage; the home team has about a 55% chance of winning. Does the current system of alternating ballparks even out the home field advantage? How often will the team that begins at home win the series?

Let's set up the simulation:

1. What is the component to be repeated?
2. How will you model each component from equally likely random digits?
3. How will you model a trial by combining components?
4. What is the response variable?
5. How will you analyze the response variable?

STEP-BY-STEP EXAMPLE Simulation

Fifty-seven students participated in a lottery for a particularly desirable dorm room—a triple with a fireplace and private bath in the tower. Twenty of the participants were members of the same varsity team. When all three winners were members of the team, the other students cried foul.

Question: Could an all-team outcome reasonably be expected to happen if everyone had a fair shot at the room?

THINK

Plan State the problem. Identify the important parts of your simulation.

I'll use a simulation to investigate whether it's unlikely that three varsity athletes would get the great room in the dorm if the lottery were fair.

Components Identify the components.

A component is the selection of a student.

Outcomes State how you will model each component using equally likely random digits. You can't just use the digits from 0 to 9 because the outcomes you are simulating are not multiples of 10%.

There are 20 and 37 students in the two groups. This time you must use *pairs* of random digits (and ignore some of them) to represent the 57 students.

I'll look at two-digit random numbers.

Let 00–19 represent the 20 varsity applicants.

Let 20–56 represent the other 37 applicants.

Skip 57–99. If I get a number in this range, I'll throw it away and go back for another two-digit random number.

Trial Explain how you will combine the components to simulate a trial. In each of these trials, you can't choose the same student twice, so you'll need to ignore a random number if it comes up a second or third time. Be sure to mention this in describing your simulation.

Each trial consists of identifying pairs of digits as V (varsity) or N (nonvarsity) until 3 people are chosen, ignoring out-of-range or repeated numbers (X)—I can't put the same person in the room twice.

Response Variable Define your response variable.

The response variable is whether or not all three selected students are on the varsity team.

SHOW

Mechanics Run several trials. Carefully record the random numbers, indicating

1) the corresponding component outcomes (here, Varsity, Nonvarsity, or ignored number) and
2) the value of the response variable.

Trial Number	Component Outcomes	All Varsity?
1	74 02 94 39 02 77 55 X V X N X X N	No
2	18 63 33 25 V X N N	No
3	05 45 88 91 56 V N X X N	No
4	39 09 07 N V V	No
5	65 39 45 95 43 X N N X N	No
6	98 95 11 68 77 12 17 X X V X X V V	Yes
7	26 19 89 93 77 27 N V X X X N	No

(continued)

8	23 52 37 N N N	No
9	16 50 83 44 V N X N	No
10	74 17 46 85 09 X V N X V	No

Analyze Summarize the results across all trials to answer the initial question.

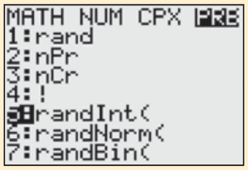
"All varsity" occurred once, or 10% of the time.

Conclusion Describe what the simulation shows, and interpret your results in the context of the real world.

In my simulation of "fair" room draws, the three people chosen were all varsity team members only 10% of the time. While this result could happen by chance, it is not particularly likely. I'm suspicious, but I'd need many more trials and a smaller frequency of the all-varsity outcome before I would make an accusation of unfairness.

TI Tips

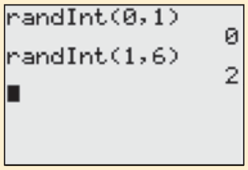
Generating random numbers



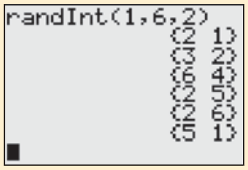
```
MATH NUM CPX PRB
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(
```

Instead of using coins, dice, cards, or tables of random numbers, you may decide to use your calculator for simulations. There are several random number generators offered in the `MATH PRB` menu.

`5:randInt(` is of particular importance. This command will produce any number of random integers in a specified range.



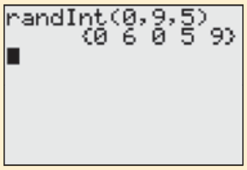
```
randInt(0,1)
             0
randInt(1,6)
             2
```



```
randInt(1,6,2)
          {2 1}
          {3 2}
          {6 4}
          {2 5}
          {2 6}
          {5 1}
```

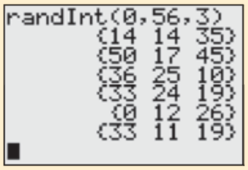
Here are some examples showing how to use `randInt` for simulations:

- `randInt(0,1)` randomly chooses a 0 or a 1. This is an effective simulation of a coin toss. You could let 0 represent tails and 1 represent heads.
- `randInt(1,6)` produces a random integer from 1 to 6, a good way to simulate rolling a die.
- `randInt(1,6,2)` simulates rolling *two* dice. To do several rolls in a row, just hit `ENTER` repeatedly.



```
randInt(0,9,5)
    {0 6 0 5 9}
```

- `randInt(0,9,5)` produces five random integers that might represent the pictures in the cereal boxes. Our run gave us two Tigers (0, 1), no Beckhams (2, 3, 4), and three Serenas (5–9).



```
randInt(0,56,3)
      {14 14 35}
      {50 17 45}
      {36 25 10}
      {33 24 19}
       {0 12 26}
      {33 11 19}
```

- `randInt(0,56,3)` produces three random integers between 0 and 56, a nice way to simulate the dorm room lottery. The window shows 6 trials, but we would skip the first one because one student was chosen twice. In none of the remaining 5 trials did three athletes (0–19) win.

WHAT CAN GO WRONG?

- **Don't overstate your case.** Let's face it: In some sense, a simulation is *always* wrong. After all, it's not the real thing. We didn't buy any cereal or run a room draw. So beware of confusing what *really* happens with what a simulation suggests *might* happen. Never forget that future results will not match your simulated results exactly.
- **Model outcome chances accurately.** A common mistake in constructing a simulation is to adopt a strategy that may appear to produce the right kind of results, but that does not accurately model the situation. For example, in our room draw, we could have gotten 0, 1, 2, or 3 team members. Why not just see how often these digits occur in random digits from 0 to 9, ignoring the digits 4 and up?

  3 2 1 7 9 0 0 5 9 7 3 7 9 2 5 2 4 1 3 8

  3 2 1 x x 0 0 x x x 3 x x 2 x 2 x 1 3 x

  This "simulation" makes it seem fairly likely that three team members would be chosen. There's a big problem with this approach, though: The digits 0, 1, 2, and 3 occur with equal frequency among random digits, making each outcome appear to happen 25% of the time. In fact, the selection of 0, 1, 2, or all 3 team members are not all equally likely outcomes. In our correct simulation, we estimated that all 3 would be chosen only about 10% of the time. If your simulation overlooks important aspects of the real situation, your model will not be accurate.
- **Run enough trials.** Simulation is cheap and fairly easy to do. Don't try to draw conclusions based on 5 or 10 trials (even though we did for illustration purposes here). We'll make precise how many trials to use in later chapters. For now, err on the side of large numbers of trials.

A S *Activity:* **Estimating Summaries from Random Outcomes.** See how well you can estimate something you can't know just by generating random outcomes.

TI-nspire

Simulations. Improve your predictions by running thousands of trials.

CONNECTIONS

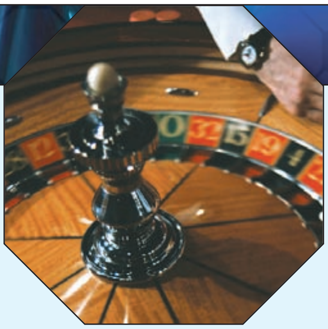
Simulations often generate many outcomes of a response variable, and we are often interested in the distribution of these responses. The tools we use to display and summarize the distribution of any real variable are appropriate for displaying and summarizing randomly generated responses as well.

Make histograms, boxplots, and Normal probability plots of the response variables from simulations, and summarize them with measures of center and spread. Be especially careful to report the variation of your response variable.

Don't forget to think about your analyses. Simulations can hide subtle errors. A careful analysis of the responses can save you from erroneous conclusions based on a faulty simulation.

You may be less likely to find an outlier in simulated responses, but if you find one, you should certainly determine how it happened.

WHAT HAVE WE LEARNED?



We've learned to harness the power of randomness. We've learned that a simulation model can help us investigate a question for which many outcomes are possible, we can't (or don't want to) collect data, and a mathematical answer is hard to calculate. We've learned how to base our simulation on random values generated by a computer, generated by a randomizing device such as a die or spinner, or found on the Internet. Like all models, simulations can provide us with useful insights about the real world.

Terms

Random 255. An outcome is random if we know the possible values it can have, but not which particular value it takes.

Generating random numbers 256. Random numbers are hard to generate. Nevertheless, several Internet sites offer an unlimited supply of equally likely random values.

Simulation 258. A simulation models a real-world situation by using random-digit outcomes to mimic the uncertainty of a response variable of interest.

Simulation component 258. A component uses equally likely random digits to model simple random occurrences whose outcomes may not be equally likely.

Trial 258. The sequence of several components representing events that we are pretending will take place.

Response variable 258. Values of the response variable record the results of each trial with respect to what we were interested in.

Skills

THINK

- Be able to recognize random outcomes in a real-world situation.
- Be able to recognize when a simulation might usefully model random behavior in the real world.

SHOW

- Know how to perform a simulation either by generating random numbers on a computer or calculator, or by using some other source of random values, such as dice, a spinner, or a table of random numbers.

TELL

- Be able to describe a simulation so that others can repeat it.
- Be able to discuss the results of a simulation study and draw conclusions about the question being investigated.

SIMULATION ON THE COMPUTER

Simulations are best done with the help of technology simply because more trials makes a better simulation, and computers are fast. There are special computer programs designed for simulation, and most statistics packages and calculators can at least generate random numbers to support a simulation.

All technology-generated random numbers are *pseudorandom.* The random numbers available on the Internet may technically be better, but the differences won't matter for any simulation of modest size. Pseudorandom numbers generate the next random value from the previous one by a specified algorithm. But they have to start somewhere. This starting point is called the "seed." Most programs let you set the seed. There's usually little reason to do this, but if you wish to, go ahead. If you reset the seed to the same value, the programs will generate the same sequence of "random" numbers.

A S ***Activity:* Creating Random Values.** Learn to use your statistics package to generate random outcomes.