

# CHAPTER 11



## Understanding Randomness

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.<sup>1</sup>

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.

---

<sup>1</sup> Don't get your hopes up.

1 2 3 4

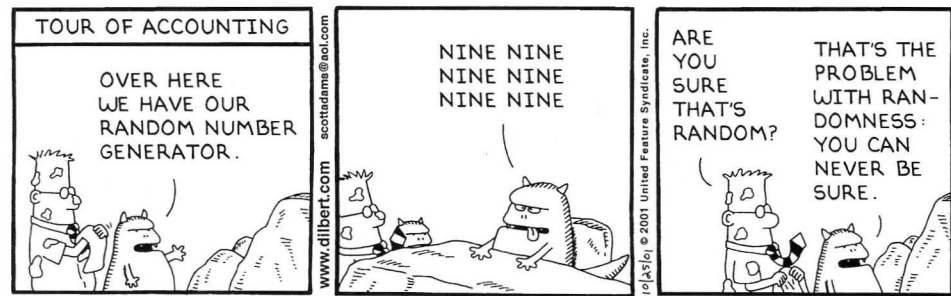
### It's Not Easy Being Random

"The generation of random numbers is too important to be left to chance."  
—Robert R. Coveyou,  
Oak Ridge National Laboratory

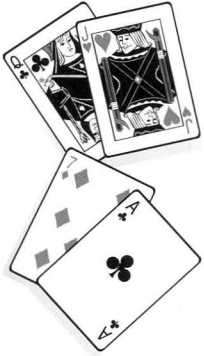
**AS** **Random Behavior.**  
ActivStats' Random Experiment Tool lets you experiment with truly random outcomes. We'll use it a lot in the coming chapters.

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.

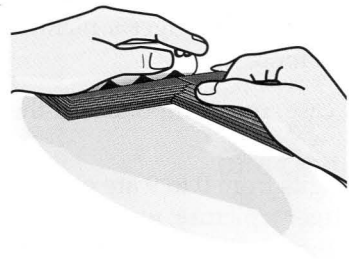
So how should we generate **random numbers**? It's surprisingly difficult to get random values even when they're equally likely. Computers have become a popular way to generate random numbers. Even though they often do much better than humans, computers can't generate truly random numbers, either. Computers 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.



**A S Truly Random Values on the Internet.** This activity takes you to an Internet site that generates all the truly random numbers you could want.



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.



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.<sup>2</sup> 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 of the decay of a radioactive element or even the random changes of lava lamps to generate truly random digits. (See this book's Web site for URLs.) In either case, a string of random digits might look like this:

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

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."

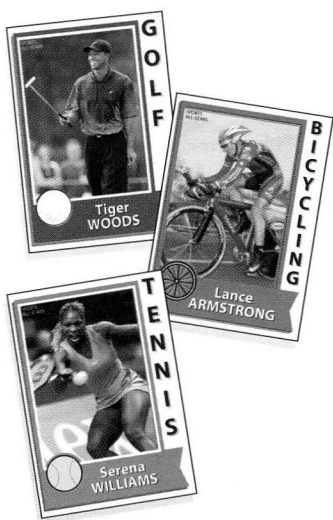
How many times should you shuffle cards to make the deck random? A surprising fact was discovered by statisticians Persi Diaconis, Ronald Graham, and W. M. Kantor. It takes seven riffle shuffles. Fewer than seven leaves order in the deck, but after that, more shuffling does little good. Most people, 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 Lance Armstrong,

<sup>2</sup> You'll find a table of random digits of this kind in the back of this book.



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 go 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 then the boxes are distributed randomly to stores around the country.

Why a model? We'll use a model to imitate cards found in random boxes of cereal. To do that, we use the random digits 0, 1, 2, . . . , 9, assuming that they are equally likely to occur and that we don't know which one will come up next. We identify 20% of the digits to represent a Tiger Woods picture, 30% to represent Lance, and the other 50% Serena. With this model we can pretend we've bought sequences of cereal boxes and simulate the outcomes—even though these outcomes (which card we get) are not equally likely—to see what happens.

## 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)

We want to use this simulation to give us some insight into how many boxes of cereal we might have to open until we get all three cards. We'll pretend to buy cereal, and the random numbers will tell us which card to pretend we got. Like any model, the simulation will be imperfect, but we hope the results we get will help us understand the situation. A **simulation** consists of a sequence of random outcomes that model a situation. In this sequence the most basic event is called a **component** of the simulation. Each component has a set of possible **outcomes**, one of which will occur at random. In our cereal example, the component is the selection of a particular box of cereal and the outcome is the type of card in the box. The sequence of events we want to investigate is called a **trial**. Trials usually involve several components; here we pretend to buy several boxes of cereal. After the trial, we record what happened—our **response variable**. By repeating this process many times—a run of the simulation—we can get an idea of what might happen if we really did try buying boxes of cereal in search of Tiger's picture. Now let's look at these steps for making a simulation in more detail.

1. **Identify the component to be repeated.** In this case, our component is the selection of a box of cereal.
2. **Explain how you will model the 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 Lance Armstrong cards, and the remaining 5 digits can represent the 50% of boxes with Serena. One possible assignment of the digits is

0, 1 = Woods      2, 3, 4 = Armstrong      5, 6, 7, 8, 9 = Williams

3. **Explain how you will simulate the trial.** A trial is the sequence of events that we are pretending will take place. In this case we want to pretend to open cereal boxes until we have one of each picture. We do this by looking at each random number and indicating what outcome it represents. We continue until we've encountered all three pictures.



For example, the random number sequences 29240 would mean you get Lance's picture (2) in the first box you open, Serena's picture (9) in the next box, two more Armstrong pictures (2, 4) in the next two boxes, and then the Tiger Woods picture (0) you need to complete your collection. Since we've gotten all three pictures, we've finished one trial of the simulation.

- 4. State clearly what the response variable is.** What are we interested in? We want to know how many boxes it takes to get all three pictures. This is the response variable. In the sample trial here, the response value is 5 boxes.
- 5. Run several trials.** A simulation is cheaper than really buying cereal, and the more trials you perform, the better. For example, consider the third line of random digits shown earlier:

89064 2730 8645681 41219 822665388587328580 1699027843110380420067...

Let's create a chart to keep track of what happened.

Trial Number	Outcomes	y = Number of boxes
1	89064 = <b>Serena</b> , Serena, <b>Tiger</b> , Serena, Lance	5
2	2730 = <b>Lance</b> , <b>Serena</b> , Lance, <b>Tiger</b>	4
3	8645681 = <b>Serena</b> , Serena, <b>Lance</b> , . . . , <b>Tiger</b>	7
4	41219 = <b>Lance</b> , <b>Tiger</b> , Lance, <b>Tiger</b> , <b>Serena</b>	5
5	822665388587328580 = <b>Serena</b> , <b>Lance</b> , . . . , <b>Tiger</b>	18

- 6. Analyze the response variable.** We wanted to know how many boxes we might expect to buy to get all three cards. To answer the question, we need to analyze the response variable in a number of ways. We know how to do this. In our first 5 trials we needed 5, 4, 7, 5, and then 18 for an average of 7.8 boxes.
- 7. State your conclusion (in the context of the problem, as always).** Based on our simulation, we estimate that customers who want the complete set of sports star pictures will buy an average of 7.8 boxes of cereal.

If you fear that this may not be an accurate estimate because we ran only five trials, you are absolutely correct. The more trials the better, and five 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.



The baseball World Series consists of up to seven games. 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. Is it an advantage to play the first two games on your home field? Or would you rather have the three middle games at home with the opportunity to win it all? In other words, do the two teams have an equal chance to win the four games needed to win the Series?

Let's set up the simulation:

- 1 What is the component to be repeated?
- 2 How will you model the outcome?
- 3 How will you simulate the trial?
- 4 What is the response variable?
- 5 How will you analyze the response variable?

**A S** **Bigger Samples Are Better.** Learn a new trick: 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.

## Simulation Step-By-Step

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. Use a simulation to determine whether an all-team outcome could 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.

**Components** Identify the components.

**Outcomes** State how you will model the random occurrence of an outcome. 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.

**Trial** Define a trial. Be sure that you don't use a repeated number this time.

**Response Variable** Define your response variable.

### Show

#### Mechanics

Run several trials. Carefully record the random numbers, indicating what each represents. Indicate the value of the response variable for each trial.

*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.*

*A component is the selection of a student.*

*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.*

*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.*

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

Trial Number	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

**Estimate**

Summarize the results across all trials.

**Tell**

**Conclusion**

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

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

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

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**

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

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

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

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

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

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.

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

### AS Estimating Summaries from Random Outcomes.

This activity throws you a curve. See how well you can estimate something you can't know (just by generating random outcomes).


## 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

<b>Random</b>	An event is random if we know what outcomes could happen, but not which particular values will happen.
<b>Random numbers</b>	Random numbers are hard to generate. Nevertheless, several Internet sites offer an unlimited supply of equally likely random values.
<b>Simulation</b>	A simulation models random events by using random numbers to specify event outcomes with relative frequencies that correspond to the true real-world relative frequencies we are trying to model.
<b>Simulation component</b>	The most basic situation in a simulation in which something happens at random.
<b>Outcome</b>	An individual result of a component of a simulation is its outcome.
<b>Trial</b>	The sequence of several components representing events that we are pretending will take place.
<b>Response variable</b>	Values of the response variable record the results of each trial with respect to what we were interested in.

## SKILLS

*When you complete this lesson you should:*

### 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 could 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, but most statistics packages can generate random numbers and 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.

### AS Creating Random Variables.

Play with your statistics package as you learn to generate random outcomes.

## EXERCISES

- Coin toss.** Is a coin flip random? Why or why not, in your opinion?
- Casino.** A casino claims that its electronic "video roulette" machine is truly random. What should that claim mean?
- The lottery.** Many states run lotteries, giving away millions of dollars if you match a certain set of winning numbers. How are those numbers determined? Do you think this method guarantees randomness? Explain.
- Games.** Many kinds of games people play rely on randomness. Cite three different methods commonly used in the attempt to achieve this randomness, and discuss the effectiveness of each.
- Bad simulations.** Explain why each of the following simulations fails to model the real situation properly.
  - Use a random integer from 0 through 9 to represent the number of heads that appear when 9 coins are tossed.
  - A basketball player takes a foul shot. Look at a random digit, using an odd digit to represent a good shot and an even digit to represent a miss.
  - Use five random digits from 1 through 13 to represent the denominations of the cards in a poker hand.
- More bad simulations.** Explain why each of the following simulations fails to model the real situation properly.
  - Use random numbers 2 through 12 to represent the sum of the faces when two dice are rolled.
  - Use a random integer from 0 through 5 to represent the number of boys in a family of 5 children.
  - Simulate a baseball player's performance at bat by letting 0 = an out, 1 = a single, 2 = a double, 3 = a triple, and 4 = a home run.
- Wrong conclusion.** A Statistics student properly simulated the length of checkout lines in a grocery store and then reported, "The average length of the line will be 3.2 people." What's wrong with this conclusion?
- Another wrong conclusion.** After simulating the spread of a disease, a researcher wrote, "24% of the people contracted the disease." What should the correct conclusion be?
- Election.** You're pretty sure that your candidate for class president has about 55% of the votes in the entire school. But you're worried that only 100 students will show up to vote. How often will the underdog (the one with 45% support) win? To find out you set up a simulation.
  - Describe how you will simulate a component and its outcomes.
  - Describe how you will simulate a trial.
  - Describe the response variable.

- 10. Two pair, or three of a kind?** When drawing five cards randomly from a deck, which is more likely, two pairs or three of a kind? A pair is exactly two of the same denomination. Three of a kind is exactly 3 of the same denomination. (Don't count three 8's as a pair—that's 3 of a kind. And don't count 4 of the same kind as two pair—that's 4 of a kind, a very special hand.) How could you simulate 5-card hands? Be careful; once you've picked the 8 of spades for a hand, you can't get it again until the next hand.
- Describe how you will simulate a component and its outcomes.
  - Describe how you will simulate a trial.
  - Describe the response variable.
- 11. Cereal.** In the chapter's example, 20% of the cereal boxes contained a picture of Tiger Woods, 30% Lance Armstrong, and the rest Serena Williams. Suppose you buy five boxes of cereal. Estimate the probability that you end up with a complete set of the pictures. Your simulation should have at least 20 runs.
- 12. Cereal, again.** Suppose you really want the Tiger Woods picture. How many boxes of cereal do you need to buy to be pretty sure of getting at least one? Your simulation should use at least 10 runs.
- 13. Multiple choice.** You take a quiz with 6 multiple choice questions. After you studied, you estimated that you would have about an 80% chance of getting any individual question right. What are your chances of getting them all right? Your simulation should use at least 20 runs.
- 14. Lucky guessing?** A friend of yours who took that same multiple choice quiz got all 6 questions right, but now claims to have guessed blindly on every question. If each question offered 4 possible answers, do you believe her? Explain, basing your argument on a simulation involving at least 10 runs.
- 15. Beat the lottery.** Many states run lotteries to raise money. A Web site advertises that it knows "how to increase YOUR chances of Winning the Lottery." They offer several systems and criticize others as foolish. One system is called *Lucky Numbers*. People who play the *Lucky Numbers* system just pick a "lucky" number to play, but maybe some numbers are luckier than others. Let's use a simulation to see how well this system works.
- To make the situation manageable, simulate a simple lottery in which a single digit from 0 to 9 is selected as the winning number. Pick a single value to bet, such as 1, and keep playing it over and over. You'll want to run at least 100 trials. (If you can program the simulations on a computer or programmable calculator, run several hundred. Or generalize the questions to a lottery that chooses two- or three-digit numbers—for which you'll need thousands of trials.)
- What proportion of the time do you expect to win?
  - Would you expect better results if you picked a "luckier" number, such as 7? (Try it, if you don't know.) Explain.
- 16. Random is as random does.** The "beat the lottery" Web site discussed in Exercise 15 suggests that because lottery numbers are random, it is better to select your bet randomly. For the same simple lottery in Exercise 15 (random values from 0 to 9), generate each bet by choosing a separate random value between 0 and 9. Play many games. What proportion of the time do you win?
- 17. It evens out in the end.** The "beat the lottery" Web site of Exercise 15 notes that in the long run we expect each value to turn up about the same number of times. That leads to their recommended betting strategy. First, watch the lottery for a while, recording all the winners. Then bet the value that has turned up the least, because we expect it will need to turn up more often to even things out. If there is more than one "rarest" value, just take the lowest one (since it doesn't matter). Simulating the simplified lottery described in Exercise 15, play many games with this system. What proportion of the time do you win?
- 18. Play the winner?** Another strategy for beating the lottery is the reverse of the system described in Exercise 17. Simulate the simplified lottery described in Exercise 15. Each time, bet the number that just turned up. The Web site suggests that this method should do worse. Does it? Play many games and see.
- 19. Driving test.** You are about to take the road test for your driver's license. You hear that only 34% of candidates pass the test the first time, but the percentage rises to 72% on subsequent retests. Estimate the average number of tests drivers take in order to get a license. Your simulation should use at least 20 runs.
- 20. Still learning?** As in Exercise 19, assume that your chance of passing the driver's test is 34% the first time and 72% for subsequent retests. Estimate the percentage of those tested who still do not have a driver's license after two attempts.
- 21. Basketball strategy.** Late in a basketball game, the team that is behind often fouls someone in an attempt to get the ball back. Usually the opposing player will get to shoot foul shots "one and one," meaning he gets a shot, and then a second shot only if he makes the first one. Suppose the opposing player has made 72% of his foul shots this season. Estimate the number of points he will score in a one-and-one situation.

- 22. Blood donors.** A person with type O-positive blood can receive blood only from other type O donors. About 44% of the U.S. population has type O blood. At a blood drive, how many potential donors do you expect to examine in order to get three units of type O blood?
- 23. Free groceries.** To attract shoppers, a supermarket runs a weekly contest that involves “scratch-off” cards. With each purchase, customers get a card with a black spot obscuring a message. When the spot is scratched away, most of the cards simply say, “Sorry—please try again.” But during the week, 100 customers will get cards that make them eligible for a drawing for free groceries. Ten of the cards say they may be worth \$200, 10 others say \$100, 20 may be worth \$50, and the rest could be worth \$20. To register those cards, customers write their names on them and put them in a barrel at the front of the store. At the end of the week the store manager draws cards at random, awarding the lucky customers free groceries in the amount specified on their card. The drawings continue until the store has given away more than \$500 of free groceries. Estimate the average number of winners each week.
- 24. Find the ace.** A new electronics store holds a contest to attract shoppers. Once an hour someone in the store is chosen at random to play the Music Game. Here’s how it works: An ace and four other cards are shuffled and placed face down on a table. The customer gets to turn cards over one at a time, looking for the ace. The person wins \$100 worth of free CDs or DVDs if the ace is the first card, \$50 if it is the second card, and \$20, \$10, or \$5 if it is the third, fourth, or fifth card chosen. What is the average dollar amount of music the store will give away?
- 25. The family.** Many couples want to have both a boy and a girl. If they decide to continue to have children until they have one child of each gender, what would the average family size be? Assume that boys and girls are equally likely.
- 26. A bigger family.** Suppose a couple will continue having children until they have at least two children of each gender (two boys *and* two girls). How many children might they expect to have?
- 27. Dice game.** You are playing a children’s game in which the number of spaces you get to move is determined by the rolling of a die. You must land exactly on the final space in order to win. If you are 10 spaces away, how many turns might it take you to win?
- 28. Parcheesi.** You are three spaces from a win in Parcheesi. On each turn you will roll two dice. To win, you must roll a total of 3 or roll a 3 on one of the dice. How many turns might you expect this to take?
- 29. The hot hand.** A basketball player with a 65% shooting percentage has just made 6 shots in a row. The announcer says this player “is hot tonight! She’s in the zone!” Assume the player takes about 20 shots per game. Is it unusual for her to make 6 or more shots in a row during a game?
- 30. The World Series.** The World Series ends when a team wins 4 games. Suppose that sports analysts consider one team a bit stronger, with a 55% chance to win any individual game. Estimate the likelihood that the underdog wins the series.
- 31. Teammates.** Four couples at a dinner party play a board game after the meal. They decide to play as teams of two and to select the teams randomly. All eight people write their names on slips of paper. The slips are thoroughly mixed, then drawn two at a time. How likely is it that every person will be teamed with someone other than the person he or she came to the party with?
- 32. Second team.** Suppose the couples in Exercise 31 choose the teams by having one member of each couple write their names on the cards and the other people each pick a card at random. How likely is it that every person will be teamed with someone other than the person he or she came with?
- 33. Job discrimination?** A company with a large sales staff announces openings for three positions as regional managers. Twenty-two of the current salespersons apply, 12 men and 10 women. After the interviews, when the company announces the newly appointed managers, all three positions go to women. The men complain of job discrimination. Do they have a case? Simulate a random selection of three people from the applicant pool and make a decision about the likelihood that a fair process would result in hiring all women.
- 34. Cell phones.** A proud legislator claims that your state’s new law against talking on a cell phone while driving has reduced cell phone use to less than 12% of all drivers. While waiting for your bus the next morning, you notice that 4 of the 10 people who drive by are using their cell phones. Does this cast doubt on the legislator’s figure of 12%? Use a simulation to estimate the likelihood of seeing at least 4 of 10 randomly selected drivers talking on their cell phones if the actual rate of usage is 12%. Explain your conclusion clearly.
- 35. Freshmen.** A certain college estimates that the 3-score SAT total for students who apply for admission can be described by a Normal model with a mean of 1570 and a standard deviation of 180. Admissions officers search the pile of envelopes, opening them at random to look for three applicants with SAT totals over 1800. How many envelopes do you think they will need to open?

# CHAPTER 12



## Sample Surveys

We have learned ways to display, describe, and summarize data. Up to now, though, our conclusions have been limited to the particular batch of data we are examining. That's OK as far as it goes, but it doesn't go very far. We usually aren't satisfied with conclusions based only on *this* group of customers, or the people who answered the survey on *this* particular day. To make business decisions, to do science, to choose wise investments, or to understand what voters think they'll do in the next election, we need to stretch beyond the data at hand to the world at large.

To make that stretch, we need three ideas. You'll find the first one natural. The second may be more surprising. The third is one of the strange but true facts that often confuse those who don't know Statistics.

### Idea 1: Examine a Part of the Whole

**AS** **Populations and Samples.** Explore the differences between populations and samples.

#### The W's and Sampling

The population we are interested in is usually determined by the *Why* of our study. The sample we draw will be the *Who*. *When* and *How* we draw the sample may depend on what is practical.

The first idea is to draw a sample. We'd like to know about an entire **population** of individuals, but examining all of them is usually impractical, if not impossible. So we settle for examining a smaller group of individuals—a **sample**—selected from the population.

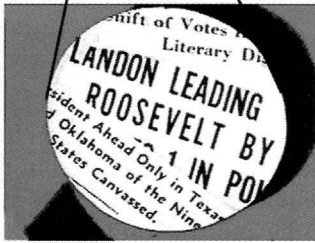
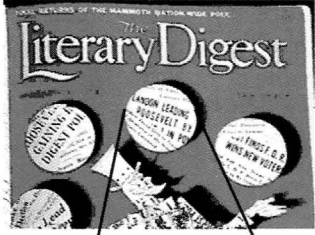
You do this every day. For example, suppose you wonder how the vegetable soup you're cooking for dinner tonight is going to go over with your friends. To decide whether it meets your standards, you only need to try a small amount. You might taste just a spoonful or two. You certainly don't have to consume the whole pot. You trust that the taste will *represent* the flavor of the entire pot. The idea behind your tasting is that a small sample, if selected properly, can represent the entire population.

It's hard to go a day without hearing about the latest opinion poll. These polls are examples of **sample surveys**, designed to ask questions of a small group of people in the hope of learning something about the entire population. Most likely, you've never been selected to be part of one of these national opinion polls. That's true of most people. So how can the pollsters claim that a sample is representative of the entire population? The answer is that professional pollsters work quite hard



to ensure that the “taste”—the sample that they take—represents the population. If not, the sample can give misleading information about the population.

## Bias



In 1936, a young pollster named George Gallup used a subsample of only 3000 of the 2.4 million responses that the *Literary Digest* received to reproduce the wrong prediction of Landon’s victory over Roosevelt. He then used an entirely different sample of 50,000 and predicted that Roosevelt would get 56% of the vote to Landon’s 44%. His sample was apparently much more *representative* of the actual voting populace. The Gallup Organization went on to become one of the leading polling companies.

Selecting a sample to represent the population fairly is more difficult than it sounds. Polls or surveys most often fail because they use a sampling method that tends to over- or underrepresent parts of the population. The method may overlook subgroups that are harder to find (such as the homeless or those who use only cell phones) or favor others (such as Internet users who like to respond to online surveys). Sampling methods that, by their nature, tend to over- or underemphasize some characteristics of the population are said to be **biased**. Bias is the bane of sampling—the one thing above all to avoid. Conclusions based on samples drawn with biased methods are inherently flawed. There is usually no way to fix bias after the sample is drawn and no way to salvage useful information from it.

What are the basic techniques for making sure that a sample is representative? Sometimes the best way to see how to do something is to study a really dismal failure. Here’s a famous one. By the beginning of the 20th century, it was common for newspapers to ask readers to return “straw” ballots on a variety of topics. (Today’s Internet surveys are the same idea, gone electronic.) The earliest known example of such a straw vote in the United States dates back to 1824.

The success of these regional polls in the early 1900s inspired national magazines to try their luck. Although the *Farm Journal* was probably the first, the *Literary Digest* was at the top of the heap. During the period 1916 to 1936, it regularly surveyed public opinion and forecast election results correctly. During the 1936 presidential campaign between Alf Landon and Franklin Delano Roosevelt, the *Literary Digest* mailed more than 10 million ballots. The magazine got back an astonishing 2.4 million. (Polls were still a relatively novel idea, and many people thought it was important to send back their opinions.) The results from the millions of responses were clear. Alf Landon would be the next president by a landslide: 57% to 43%. You remember President Landon, don’t you? In fact, Landon carried only two states. Roosevelt won, 62% to 37%, and, perhaps coincidentally, the *Digest* went bankrupt soon afterward.

What went wrong? The problem was that the *Digest* sample was not representative. The pollsters made some mistakes that are now considered classics. First, let’s look at how they got the list of 10 million names to start with. Where would you go to get such a list? You might think of using phone numbers as a way to select people—and that’s just what the *Digest* did. But in 1936, at the height of the Great Depression, telephones were real luxuries. Any list of phone owners would include far more rich than poor people. In fact, it wasn’t until 1986 that enough families in the United States had telephones so that phoning became a reliable way of surveying people.<sup>1</sup> The other lists available to the *Digest* were even less representative—drivers’ registrations and memberships in organizations such as country clubs.

<sup>1</sup> Even today, phone numbers must be computer-generated to make sure that the phone owners are representative. Using only phone book listings would miss people with unlisted numbers, cell phones, or who have recently moved. Leaving these groups out may make the sample unrepresentative.



**AS** **The Literary Digest Poll and the Election of 1936.** Watch the story of one of the most famous polling failures in history. (Turn ahead in the Lesson Book to find this one.)

The main campaign issue in 1936 was the economy. Roosevelt's core supporters, who tended to be less well off, were not well represented in the *Digest's* sample, so the results of a survey based on that sample did not reflect the opinions of the overall population. It did not matter how well the sample was measured, nor how many people responded.

How can we avoid the *Digest's* errors? To avoid bias and make the sample as representative as possible, you might be tempted to handpick the individuals included in the sample with care and precision. The best strategy is to do something quite different. We should select individuals for the sample *at random*. The value of deliberately introducing randomness is one of the great insights of Statistics.

## Idea 2: Randomize



Think back to the soup sample. Suppose you add some salt to the pot. If you sample it from the top before stirring, what will happen? With the salt sitting on top, you'll get the misleading idea that the whole pot is salty. If you sample from the bottom, you'll get an equally misleading idea that the whole pot is bland. By stirring, you *randomize* the amount of salt throughout the pot, making each taste more typical in terms of the amount of salt in the whole pot.

Randomization can protect you against factors that you know are in the data. It can also help protect against factors that you aren't even aware of. Suppose, while you weren't looking, a friend added a handful of peas to the soup. They are down at the bottom of the pot, mixing with the other vegetables. If you don't randomize the soup by stirring, your test spoonful from the top won't have any peas. By stirring in the salt, you *also* randomize the peas throughout the pot, making your sample taste more typical of the overall pot *even though you didn't know the peas were there*. So randomizing protects us by giving us a representative sample even over effects we were unaware of.

How would we "stir" people in our survey? We'd try to select them at random. **Randomizing** protects us from the influences of *all* the features of our population, even ones that we may not have thought about. It does that by making sure that *on average* the sample looks like the rest of the population.

● **Why not match the sample to the population?** Rather than randomizing, we could try to design our sample so that the people we choose are typical in terms of every characteristic we can think of. In the 1936 vote, rich and poor voted differently as in no previous election. So, we'd like the income levels of those we sample to match the population. How about age? Do young and old vote alike? Political affiliation? Marital status? Having children? Living in the suburbs? We can't possibly think of all the things that might be important. Even if we could, we wouldn't be able to match our sample to the population for all these characteristics. ●

**AS** **Sampling from Some Real Populations.** Draw random samples for yourself to see how closely they resemble each other and the population.

Not only does randomizing protect us from bias, it actually makes it possible for us to draw inferences about the population when we see only a sample. Such inferences are among the most powerful things we can do with Statistics, and we'll spend much of the rest of the book discussing them. Keep in mind, though, that it's all made possible because we deliberately choose things randomly.

**NOTATION ALERT:**

This entire table is a notation alert.

Name	Statistic	Parameter
Mean	$\bar{y}$	$\mu$ (mu, pronounced “meeoo,” not “moo”)
Standard deviation	$s$	$\sigma$ (sigma)
Correlation	$r$	$r$ (rho)
Regression coefficient	$b$	$\beta$ (beta, pronounced “baytah” <sup>5</sup> )
Proportion	$\hat{p}$	$p$ (pronounced “pee” <sup>6</sup> )



- 1 Various claims are often made for surveys. Why is each of the following claims not correct?
- It is always better to take a census than to draw a sample.
  - Stopping students on their way out of the cafeteria is a good way to sample if we want to know about the quality of the food there.
  - We drew a sample of 100 from the 3000 students in a school. To get the same level of precision for a town of 30,000 residents, we'll need a sample of 1000.
  - A poll taken at our favorite Web site ([www.statsisfun.org](http://www.statsisfun.org)) garnered 12,357 responses. The majority said they enjoy doing statistics homework. With a sample size that large, we can be pretty sure that most Statistics students feel this way, too.
  - The true percentage of all Statistics students who enjoy the homework is called a “population statistic.”

## Simple Random Samples

We draw samples because we can't work with the entire population. We need to be sure that the statistics we compute from the sample reflect the corresponding parameters accurately. A sample that does this is said to be **representative**.

How would you select a representative sample? Most people would say that every individual in the population should have an equal chance to be selected, and certainly that seems fair. But it's not sufficient. There are many ways to give everyone an equal chance that still wouldn't give a representative sample. Consider, for example, a school that has equal numbers of males and females. We could sample like this. Flip a coin. If it comes up heads, select 100 female students at random. If it comes up tails, select 100 males at random. Everyone has an equal chance of selection, but every sample is of only a single sex—hardly representative.

We need to do better. Suppose we insist that every possible *sample* of the size we plan to draw has an equal chance to be selected. This ensures that situations like the one just described are not likely to occur and still guarantees that each person has an equal chance of being selected. What's different is that with this method each *combination* of people has an equal chance of being selected as well. A sample drawn in this way is called a **Simple Random Sample**, usually abbreviated **SRS**. An SRS is the standard against which we measure other sampling methods, and the sampling method on which the theory of working with sampled data is based.

<sup>5</sup> If you're American. If you're British or Canadian, it's “beetah.”

<sup>6</sup> Just in case you weren't sure.

To select a sample at random, we first need to define where the sample will come from. The **sampling frame** is a list of individuals from which the sample is drawn. For example, to draw a random sample of students at a college, we might obtain a list of all registered full-time students and sample from that list. In defining the sampling frame, we must deal with the details of defining the population. Are part-time students included? How about those who are attending school elsewhere and transferring credits back to the college?

Once we have a sampling frame, the easiest way to choose an SRS is with random numbers. We already talked about some good ways to get random numbers in Chapter 11. We can assign a random number to each individual in the sampling frame. As in a simulation, we then select only those whose random numbers satisfy some rule. Let's look at a couple of examples:

- We want to select 5 students from the 80 enrolled in an Introductory Statistics class. We start by numbering the students from 00 to 79. Now we get a sequence of random digits from a table, technology, or the Internet: 05166 29305 77482. Taking those random numbers two digits at a time gives us 05, 16, 62, 93, 05, 77, and 48. We ignore 93 because no one had a number that high. And, so as not to pick the same person twice, we also skip the repeated number 05. Our simple random sample consists of students with the numbers 05, 16, 62, 77, and 48.
- Often the sampling frame is so large that it would be too tedious to number everyone consecutively. If our intended sample size is approximately 10% of the sampling frame, we assign each individual a single random digit 0 to 9. Then we select only those with a specific random digit, say, 5.

Samples drawn at random generally differ one from another. Each draw of random numbers selects *different* people for our sample. These differences lead to different values for the variables we measure. We call these sample-to-sample differences **sampling variability**. Surprisingly, sampling variability isn't a problem; it's an opportunity. If different samples from a population vary little from each other, then most likely the underlying population harbors little variation. If the samples show much sampling variability, the underlying population probably varies a lot. In the coming chapters, we'll spend much time and attention working with sampling variability to better understand what we are trying to measure.

#### Error Okay, Bias Bad!

Sampling variability is sometimes referred to as *sampling error*, making it sound like it's some kind of mistake. It's not. We understand that samples will vary, so "sampling error" is to be expected. It's *bias* we must strive to avoid. Bias means our sampling method distorts our view of the population, and that will surely lead to mistakes.

## Stratified Sampling

Simple random sampling is not the only fair way to sample. More complicated designs may save time or money or help avoid sampling problems. All statistical sampling designs have in common the idea that chance, rather than human choice, is used to select the sample.

Designs that are used to sample from large populations—especially populations residing across large areas—are often more complicated than simple random samples. Sometimes the population is first sliced into homogeneous groups, called **strata**, before the sample is selected. Then simple random sampling is used within each stratum before the results are combined. This common sampling design is called **stratified random sampling**.

Why would we want to complicate things? Here's an example. Suppose we want to survey how students feel about funding for the football team at a large university. The campus is 60% men and 40% women, and we suspect that men and women have different views on the funding. If we use simple random sampling to

select 100 people for the survey, we could end up with 70 men and 30 women or 35 men and 65 women. Our resulting estimates of the level of support for the football funding could vary widely. To help reduce this sampling variability, we can decide to force a representative gender balance, selecting 60 men at random and 40 women at random. This would guarantee that the proportions of men and women within our sample match the proportions in the population, and that should make such samples more accurate in representing population opinion.

You can imagine the importance of stratifying by race, income, age, and other characteristics, depending on the questions in the survey. When we use a sampling method that restricts by strata, additional samples are more like one another, so statistics calculated for the sampled values will vary less from one sample to another. This reduced sampling variability is the most important benefit of stratifying.

Stratified sampling can also help us notice important differences among groups. As we saw in Chapter 3, if we unthinkingly combine group data we risk reaching the wrong conclusion, becoming victims of Simpson's paradox.

## Cluster and Multistage Sampling

Sometimes dividing the sample into homogeneous strata isn't practical. And even simple random sampling may be difficult. For example, suppose we wanted to assess the reading level of this textbook based on the length of the sentences. Simple random sampling could be awkward; we'd have to number each sentence, and then find, for example, the 576th sentence or the 2482nd sentence, and so on. Doesn't sound like much fun, does it?

We could make our task much easier by picking a few *pages* at random and then counting the lengths of the sentences on those pages. That's easier than picking individual sentences, and works if we believe that the pages are all reasonably similar to one another in terms of reading level. Splitting the population into similar parts or **clusters** can make sampling more practical. Then we could simply select one or a few clusters at random and perform a census within each of them. This sampling design is called **cluster sampling**. If each cluster fairly represents the full population, cluster sampling will give us an unbiased sample.

What's the difference between cluster sampling and stratified sampling? We stratify to ensure that our sample represents different groups in the population, and sample randomly within each stratum. Strata are homogeneous, but differ from one another. By contrast, clusters are more or less alike, each heterogeneous and resembling the overall population. We select clusters to make sampling more practical or affordable.

Sometimes we use a variety of sampling methods together. In trying to assess the reading level of this book, we might worry that it starts out easy and then gets harder as the concepts become more difficult. If so, we'd want to avoid samples that selected heavily from early or from late chapters. To guarantee a fair mix of chapters, we could randomly choose one chapter from each of the seven parts of the book. Then we would randomly select a few pages from each of those chapters. If altogether that made too many sentences, we might select a few sentences at random from each of the chosen pages. So, what is our sampling strategy? First we stratify by the part of the book, and randomly choose a chapter to represent each stratum. Within each selected chapter, we choose pages as clusters. Finally, we consider an SRS of sentences within each cluster. Sampling schemes that combine several methods are called **multistage samples**. Most surveys conducted by

### Strata, or Clusters?

We may split a population into strata or clusters. What's the difference? We create strata by dividing the population into groups of similar individuals so that each stratum is different from the others. (For example, we often stratify by age, race, or sex.) By contrast, we create clusters that all look pretty much alike, each representing the wide variety of individuals seen in the population.

professional polling organizations use some combination of stratified and cluster sampling as well as simple random samples.

## Systematic Samples

Sometimes we draw a sample by selecting individuals systematically. For example, you might survey every 10th person on an alphabetical list of students. To make it random, you still must start the systematic selection from a randomly selected individual. When there is no reason to believe that the order of the list could be associated in any way with the responses sought, **systematic sampling** can give a representative sample. Systematic sampling can be much less expensive than true random sampling. When you use a systematic sample, you should justify the assumption that the systematic method is not associated with any of the measured variables.

Think about the reading level sampling example again. Suppose we have chosen a chapter of the book at random, then three pages at random from that chapter, and now we want to select a sample of 10 sentences from the 73 sentences found on those pages. Instead of numbering each sentence so we can pick a simple random sample, it would be easier to sample systematically. A quick calculation shows  $73/10 = 7.3$ , so we can get our sample by just picking every seventh sentence on the page. But where should you start? At random, of course. We've accounted for  $10 \times 7 = 70$  of the sentences, so we'll throw the extra 3 into the starting group and choose a sentence at random from the first 10. Then we pick every seventh sentence after that and record its length.



- 2 We need to survey a random sample of the 300 passengers on a flight from San Francisco to Tokyo. Name each sampling method described below.
- Pick every 10th passenger as people board the plane.
  - From the boarding list, randomly choose 5 people flying first class and 25 of the other passengers.
  - Randomly generate 30 seat numbers and survey the passengers who sit there.
  - Randomly select a seat position (right window, right center, right aisle, etc.) and survey all the passengers sitting in those seats.

## Sampling Step-By-Step

The assignment says, "Conduct your own sample survey to find out how many hours per week students at your school spend watching TV during the school year." Let's see how we might do this step by step. (Remember, though—actually collecting the data from your sample can be difficult and time consuming.)

### Think

**Plan** State what you want to know.

### Population and Parameter

Identify the *W*'s of the study. The *Why* determines the population and the associated sampling frame. The *What* identifies the parameter of interest and the variables measured.

*I wanted to design a study to find out how many hours of TV students at my school watch.*

*The population studied was students at our school. I obtained a list of all students currently enrolled and used it as the sampling frame. The parameter of interest was the number of TV hours watched per week during the*



## TERMS

**Population** The entire group of individuals or instances about whom we hope to learn.

**Sample** A (representative) subset of a population, examined in hope of learning about the population.

**Sample survey** A study that asks questions of a sample drawn from some population in the hope of learning something about the entire population. Polls taken to assess voter preferences are common sample surveys.

**Bias** Any systematic failure of a sampling method to represent its population is bias. It is almost impossible to recover from bias, so efforts to avoid it are well spent. Common errors include

- relying on voluntary response.
- undercoverage of the population.
- nonresponse bias.
- response bias.

**Randomization** The best defense against bias is randomization, in which each individual is given a fair, random chance of selection.

**Matching** Any attempt to force a sample to resemble specified attributes of the population is a form of matching. Matching may help make better samples, but it is no substitute for randomizing.

**Sample size** The number of individuals in a sample. The sample size determines how well the sample represents the population, not the fraction of the population sampled.

**Census** A sample that consists of the entire population is called a census.

**Population parameter** A numerically valued attribute of a model for a population. We rarely expect to know the true value of a population parameter, but we do hope to estimate it from sampled data. For example, the mean income of all employed people in the country is a population parameter.

**Statistic, sample statistic** Statistics are values calculated for sampled data. Those that correspond to, and thus estimate, a population parameter, are of particular interest. For example, the mean income of all employed people in a representative sample can provide a good estimate of the corresponding population parameter. The term “sample statistic” is sometimes used, usually to parallel the corresponding term, “population parameter.”

**Representative** A sample is said to be representative if the statistics computed from it accurately reflect the corresponding population parameters.

**Simple random sample (SRS)** A simple random sample of sample size  $n$  is one in which each set of  $n$  elements in the population has an equal chance of selection.

**Sampling frame** A list of individuals from whom the sample is drawn is called the sampling frame. Individuals who may be in the population of interest but who are not in the sampling frame cannot be included in any sample.

**Sampling variability** The natural tendency of randomly drawn samples to differ, one from another. Sometimes, unfortunately, called *sampling error*, sampling variability is no error at all, but just the natural result of random sampling.

**Stratified random sample** A sampling design in which the population is divided into several subpopulations, or **strata**, and random samples are then drawn from each stratum. If the strata are homogeneous but are different from each other, a stratified sample may yield more consistent results.

<b>Cluster sample</b>	A sampling design in which entire groups, or <b>clusters</b> , are chosen at random. Cluster sampling is usually selected as a matter of convenience, practicality, or cost. Each cluster should be heterogeneous (and representative of the population), so all the clusters should be similar to each other.
<b>Multistage sample</b>	Sampling schemes that combine several sampling methods are called multistage samples. For example, a national polling service may stratify the country by geographical regions, select a random sample of cities from each region, and then interview a cluster of residents in each city.
<b>Systematic sample</b>	A sample drawn by selecting individuals systematically from a sampling frame. When there is no relationship between the order of the sampling frame and the variables of interest, a systematic sample can be representative.
<b>Voluntary response bias</b>	Bias introduced to a sample when individuals can choose on their own whether to participate in the sample. Samples based on voluntary response are always invalid and cannot be recovered, no matter how large the sample size.
<b>Convenience sample</b>	A convenience sample consists of the individuals who are conveniently available. Convenience samples often fail to be representative because every individual in the population is not equally convenient to sample.
<b>Undercoverage</b>	A sampling scheme that biases the sample in a way that gives a part of the population less representation than it has in the population, suffers from undercoverage.
<b>Nonresponse bias</b>	Bias introduced to a sample when a large fraction of those sampled fails to respond. Those who do respond are likely to not represent the entire sample. Voluntary response bias is a form of nonresponse bias, but nonresponse may occur for other reasons. For example, those who are at work during the day won't respond to a telephone survey conducted only during working hours.
<b>Response bias</b>	Anything in a survey design that influences responses falls under the heading of response bias. One typical response bias arises from the wording of questions, which may suggest a favored response. Voters, for example, are more likely to express support of "the president" than support of the particular person holding that office at the moment.

## SKILLS

*When you complete this lesson you should:*

### Think

- Know the basic concepts and terminology of sampling (see the preceding list).
- Recognize population parameters in descriptions of populations and samples.
- Understand the value of randomization as a defense against bias.
- Understand the value of sampling to estimate population parameters from statistics calculated on representative samples drawn from the population.
- Understand that the size of the sample (not the fraction of the population) determines the precision of estimates.

### Show

- Know how to draw a simple random sample from a master list of a population, using a computer or a table of random numbers.

### Tell

- Know what to report about a sample as part of your account of a statistical analysis.
- Report possible sources of bias in sampling methods. Recognize voluntary response and nonresponse as sources of bias in a sample survey.

## Sampling on the Computer

Computer-generated pseudorandom numbers are usually quite good enough for drawing random samples. But there is little reason not to use the truly random values available on the Internet.

Here's a convenient way to draw an SRS of a specified size using a computer-based sampling frame. The sampling frame can be a list of names or of identification numbers arrayed, for example, as a column in a spreadsheet, statistics program, or database:

1. Generate random numbers of enough digits so that each exceeds the size of the sampling frame list by several digits. This makes duplication unlikely.
2. Assign the random numbers arbitrarily to individuals in the sampling frame list. For example, put them in an adjacent column.
3. Sort the list of random numbers, *carrying* along the sampling frame list.
4. Now the first  $n$  values in the sorted sampling frame column are an SRS of  $n$  values from the entire sampling frame.

## EXERCISES

1–10. *What did they do? For the following reports about statistical studies, identify the following items (if possible). If you can't tell, then say so—this often happens when we read about a survey.*

- a) The population
  - b) The population parameter of interest
  - c) The sampling frame
  - d) The sample
  - e) The sampling method, including whether or not randomization was employed
  - f) Any potential sources of bias you can detect and any problems you see in generalizing to the population of interest
1. A business magazine mailed a questionnaire to the human resource directors of all of the Fortune 500 companies, and received responses from 23% of them. Those responding reported that they did not find that such surveys intruded significantly on their workday.
  2. A question posted on the Lycos Web site on 18 June 2000 asked visitors to the site to say whether they thought that marijuana should be legally available for medicinal purposes.
  3. Consumers Union asked all subscribers whether they had used alternative medical treatments and, if so, whether they had benefited from them. For almost all of the treatments, approximately 20% of those responding reported cures or substantial improvement in their condition.
  4. The Gallup Poll interviewed 1423 randomly selected American citizens September 10–14, 1999, and reported that when “asked which type of content bothers them most on TV, 44% of Americans identify ‘violence,’ 23% choose ‘lewd and profane language,’ while 22% say ‘sexual situations.’”
  5. Researchers waited outside a bar they had randomly selected from a list of such establishments. They stopped every 10th person who came out of the bar and asked whether he or she thought drinking and driving was a serious problem.
  6. Hoping to learn what issues may resonate with voters in the coming election, the campaign director for a mayoral candidate selects one block from each of the city's election districts. Staff members go there and interview all the residents they can find.
  7. The Environmental Protection Agency took soil samples at 16 locations near a former industrial waste dump and checked each for evidence of toxic chemicals. They found no elevated levels of any harmful substances.
  8. State police set up a roadblock to check cars for up-to-date registration, insurance, and safety inspections. They usually find problems with about 10% of the cars they stop.
  9. A company packaging snack foods maintains quality control by randomly selecting 10 cases from each day's production and weighing the bags. Then they open one bag from each case and inspect the contents.

- 10. Dairy inspectors** visit farms unannounced and take samples of the milk to test for contamination. If the milk is found to contain dirt, antibiotics, or other foreign matter, the milk will be destroyed and the farm re-inspected until purity is restored.
- 11. Parent opinion, part 1.** In a large city school system with 20 elementary schools, the school board is considering the adoption of a new policy that would require elementary students to pass a test in order to be promoted to the next grade. The PTA wants to find out whether parents agree with this plan. Listed below are some of the ideas proposed for gathering data. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
- Put a big ad in the newspaper asking people to log their opinions on the PTA Web site.
  - Randomly select one of the elementary schools and contact every parent by phone.
  - Send a survey home with every student, and ask parents to fill it out and return it the next day.
  - Randomly select 20 parents from each elementary school. Send them a survey, and follow up with a phone call if they do not return the survey within a week.
- 12. Parent opinion, part 2.** Let's revisit the school system described in Exercise 11. Four new sampling strategies have been proposed to help the PTA determine whether parents favor requiring elementary students to pass a test in order to be promoted to the next grade. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
- Run a poll on the local TV news, asking people to dial one of two phone numbers to indicate whether they favor or oppose the plan.
  - Hold a PTA meeting at each of the 20 elementary schools and tally the opinions expressed by those who attend the meetings.
  - Randomly select one class at each elementary school and contact each of those parents.
  - Go through the district's enrollment records, selecting every 40th parent. PTA volunteers will go to those homes to interview the people chosen.
- 13. Churches.** For your political science class, you'd like to take a survey from a sample of all the Catholic Church members in your city. A list of churches shows 17 Catholic churches within the city limits. Rather than try to obtain a list of all members of all these churches, you decide to pick 3 churches at random. For those churches, you'll ask to get a list of all current members and contact 100 members at random.
- What kind of design have you used?
  - What could go wrong with the design that you have proposed?
- 14. Fish.** The U.S. Fish and Wildlife Service plans to study the kinds of fish being taken out of Saginaw Bay. To do that, they decide to randomly select 5 fishing boats at the end of a randomly chosen fishing day and to count the numbers and types of all the fish on those boats.
- What kind of design have they used?
  - What could go wrong with the design that they have proposed?
- 15. Roller coasters.** An amusement park has opened a new roller coaster. It is so popular that people are waiting for up to 3 hours for a 2-minute ride. Concerned about how patrons (who paid a large amount to enter the park and ride on the rides) feel about this, they survey every 10th person on the line for the roller coaster, starting from a randomly selected individual.
- What kind of sample is this?
  - Is it likely to be representative?
  - What is the sampling frame?
- 16. Playground.** Some people have been complaining that the children's playground at a municipal park is too small and is in need of repair. Managers of the park decide to survey city residents to see if they believe the playground should be rebuilt. They hand out questionnaires to parents who bring children to the park. Describe possible biases in this sample.
- 17. Wording the survey.** Two members of the PTA committee in Exercises 11 and 12 have proposed different questions to ask in seeking parents' opinions.
- Question 1: Should elementary school-age children have to pass high stakes tests in order to remain with their classmates?*
- Question 2: Should schools and students be held accountable for meeting yearly learning goals by testing students before they advance to the next grade?*
- Do you think responses to these two questions might differ? How? What kind of bias is this?
  - Propose a question with more neutral wording that might better assess parental opinion.
- 18. Banning ephedra.** An online poll at a popular Web site asked:
- A nationwide ban of the diet supplement ephedra went into effect recently. The herbal stimulant has been linked to 155 deaths and many more heart attacks and strokes. Ephedra manufacturer NVE Pharmaceuticals, claiming that the FDA lacked proof that ephedra is dangerous if used as directed, was denied a temporary restraining order on the ban yesterday by a federal judge. Do you think that ephedra should continue to be banned nationwide?*
- 65% of 17,303 respondents said "yes." Comment on each of the following statements about this poll:
- With a sample size that large, we can be pretty certain we know the true proportion of Americans who think ephedra should be banned.
  - The wording of the question is clearly very biased.

- c) The sampling frame is all Internet users.
- d) This is a voluntary response survey, so the results can't be reliably generalized to any population of interest.

**19. Another ride.** The survey of patrons waiting in line for the roller coaster in Exercise 15 asks whether they think it is worthwhile to wait a long time for the ride and whether they'd like the amusement park to install still more roller coasters. What biases might cause a problem for this survey?

**20. Playground, act two.** The survey described in Exercise 16 asked:

*Many people believe this playground is too small and in need of repair. Do you think the playground should be repaired and expanded even if that means raising the entrance fee to the park?*

Describe two ways this question may lead to response bias.

**21. Survey questions.** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.

- a) Should companies that pollute the environment be compelled to pay the costs of cleanup?
- b) Given that 18-year-olds are old enough to vote and to serve in the military, is it fair to set the drinking age at 21?

**22. More survey questions.** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.

- a) Do you think high-school students should be required to wear uniforms?
- b) Given humanity's great tradition of exploration, do you favor continued funding for space flights?

**23. Phone surveys.** Anytime we conduct a survey we must take care to avoid undercoverage. Suppose we plan to select 500 names from the city phone book, call their homes between noon and 4 p.m., and interview whoever answers, anticipating contacts with at least 200 people.

- a) Why is it difficult to use a simple random sample here?
- b) Describe a more convenient, but still random, sampling strategy.
- c) What kinds of households are likely to be included in the eventual sample of opinion? Who will be excluded?
- d) Suppose, instead, that we continue calling each number, perhaps in the morning or evening, until an adult is contacted and interviewed. How does this improve the sampling design?
- e) Random digit dialing machines can generate the phone calls for us. How would this improve our design? Is anyone still excluded?

**24. Cell phone survey.** What about drawing a random sample only from cell phone exchanges? Discuss the advantages and disadvantages of such a sampling method as compared with surveying randomly generated telephone numbers from non-cell phone exchanges. Do you think these advantages and disadvantages have changed over time? How do you expect they'll change in the future?

tages and disadvantages of such a sampling method as compared with surveying randomly generated telephone numbers from non-cell phone exchanges. Do you think these advantages and disadvantages have changed over time? How do you expect they'll change in the future?

**25. Arm length.** How long is your arm compared with your hand size? Put your right thumb at your left shoulder bone, stretch your hand open wide, and extend your hand down your arm. Put your thumb at the place where your little finger is and extend down the arm again. Repeat this a third time. Now your little finger will probably have reached the back of your left hand. If the fourth hand width goes past the end of your middle finger, turn your hand sideways and count finger widths to get there.

- a) How many hand and finger widths is your arm?
- b) Suppose you repeat your measurement 10 times, and average your results. What parameter would this average estimate? What is the population?
- c) Suppose you now collect arm lengths measured in this way from 9 friends and average these 10 measurements. What is the population now? What parameter would this average estimate?
- d) Do you think these 10 arm lengths are likely to be representative of the population of arm lengths in your community? In the country? Why or why not?

**26. Fuel economy.** Occasionally when I fill my car with gas, I figure out how many miles per gallon my car got. I wrote down those results after 6 fill-ups in the past few months. Overall, it appears my car gets 28.8 miles per gallon.

- a) What statistic have I calculated?
- b) What is the parameter I'm trying to estimate?
- c) How might my results be biased?
- d) When the Environmental Protection Agency (EPA) checks a car like mine to predict its fuel economy, what parameter is it trying to estimate?

**27. Accounting.** Between quarterly audits, a company likes to check on its accounting procedures to address any problems before they become serious. The accounting staff processes payments on about 120 orders each day. The next day, the supervisor rechecks 10 of the transactions to be sure they were processed properly.

- a) Propose a sampling strategy for the supervisor.
- b) How would you modify that strategy if the company makes both wholesale and retail sales, requiring different bookkeeping procedures?

**28. Happy workers?** A manufacturing company employs 14 project managers, 48 foremen, and 377 laborers. In an effort to keep informed about any possible sources of employee discontent, management wants to conduct job satisfaction interviews with a sample of employees every month.

- a) Do you see any danger of bias in the company's plan? Explain.



- b) Propose a sampling strategy that uses a simple random sample.
- c) Why do you think a simple random sample might not provide the representative opinion the company seeks?
- d) Propose a better sampling strategy.
- e) Listed below are the last names of the project managers. Use random numbers to select two people to be interviewed. Be sure to explain your method carefully.

Barrett	Bowman	Chen
DeLara	DeRoos	Grigorov
Maceli	Mulvaney	Pagliarulo
Rosica	Smithson	Tadros
Williams	Yamamoto	

- 29. Quality control.** Sammy's Salsa, a small local company, produces 20 cases of salsa a day. Each case contains 12 jars and is imprinted with a code indicating the date and batch number. To help maintain consistency, at the end of each day Sammy selects three bottles of salsa, weighs the contents, and tastes the product. Help Sammy select the sample jars. Today's cases are coded 07N61 through 07N80.
- a) Carefully explain your sampling strategy.
  - b) Show how to use random numbers to pick the three jars for testing.
  - c) Did you use a simple random sample? Explain.

- 30. A fish story.** Concerned about reports of discolored scales on fish caught downstream from a newly sited chemical plant, scientists set up a field station in a shoreline public park. For one week they asked fishermen there to bring any fish they caught to the field station for a brief inspection. At the end of the week, the scientists said that 18% of the 234 fish that were submitted for inspection displayed the discoloration. From this information, can the researchers estimate what proportion of fish in the river have discolored scales? Explain.

- 31. Sampling methods.** Consider each of these situations. Do you think the proposed sampling method is appropriate? Explain.

- a) We want to know what percentage of local doctors accept Medicaid patients. We call the offices of 50 doctors randomly selected from local Yellow Page listings.
- b) We want to know what percentage of local businesses anticipate hiring additional employees in the upcoming month. We randomly select a page in the Yellow Pages, and call every business listed there.

- 32. More sampling methods.** Consider each of these situations. Do you think the proposed sampling method is appropriate? Explain.

- a) We want to know if there is neighborhood support to turn a vacant lot into a playground. We spend a Saturday afternoon going door-to-door in the neighborhood, asking people to sign a petition.
- b) We want to know if students at our college are satisfied with the selection of food available on campus. We go to the largest cafeteria and interview every 10th person in line.

### just checking

#### Answers

1.
  - a) It can be hard to reach all members of a population, and it can take so long that circumstances change, affecting the responses. A well-designed sample is often a better choice.
  - b) This sample is probably biased—students who didn't like the food at the cafeteria might not choose to eat there.
  - c) No, only the sample size matters, not the fraction of the overall population.
  - d) Students who frequent this Web site might be more enthusiastic about Statistics than the overall population of Statistics students. A large sample cannot compensate for bias.
  - e) It's the population "parameter." "Statistics" describe samples.
2.
  - a) systematic
  - b) stratified
  - c) simple
  - d) cluster

# A P P E N D I X **G** Tables

Row	TABLE OF RANDOM DIGITS									
1	96299	07196	98642	20639	23185	56282	69929	14125	38872	94168
2	71622	35940	81807	59225	18192	08710	80777	84395	69563	86280
3	03272	41230	81739	74797	70406	18564	69273	72532	78340	36699
4	46376	58596	14365	63685	56555	42974	72944	96463	63533	24152
5	47352	42853	42903	97504	56655	70355	88606	61406	38757	70657
6	20064	04266	74017	79319	70170	96572	08523	56025	89077	57678
7	73184	95907	05179	51002	83374	52297	07769	99792	78365	93487
8	72753	36216	07230	35793	71907	65571	66784	25548	91861	15725
9	03939	30763	06138	80062	02537	23561	93136	61260	77935	93159
10	75998	37203	07959	38264	78120	77525	86481	54986	33042	70648
11	94435	97441	90998	25104	49761	14967	70724	67030	53887	81293
12	04362	40989	69167	38894	00172	02999	97377	33305	60782	29810
13	89059	43528	10547	40115	82234	86902	04121	83889	76208	31076
14	87736	04666	75145	49175	76754	07884	92564	80793	22573	67902
15	76488	88899	15860	07370	13431	84041	69202	18912	83173	11983
16	36460	53772	66634	25045	79007	78518	73580	14191	50353	32064
17	13205	69237	21820	20952	16635	58867	97650	82983	64865	93298
18	51242	12215	90739	36812	00436	31609	80333	96606	30430	31803
19	67819	00354	91439	91073	49258	15992	41277	75111	67496	68430
20	09875	08990	27656	15871	23637	00952	97818	64234	50199	05715
21	18192	95308	72975	01191	29958	09275	89141	19558	50524	32041
22	02763	33701	66188	50226	35813	72951	11638	01876	93664	37001
23	13349	46328	01856	29935	80563	03742	49470	67749	08578	21956
24	69238	92878	80067	80807	45096	22936	64325	19265	37755	69794
25	92207	63527	59398	29818	24789	94309	88380	57000	50171	17891
26	66679	99100	37072	30593	29665	84286	44458	60180	81451	58273
27	31087	42430	60322	34765	15757	53300	97392	98035	05228	68970
28	84432	04916	52949	78533	31666	62350	20584	56367	19701	60584
29	72042	12287	21081	48426	44321	58765	41760	43304	13399	02043
30	94534	73559	82135	70260	87936	85162	11937	18263	54138	69564
31	63971	97198	40974	45301	60177	35604	21580	68107	25184	42810
32	11227	58474	17272	37619	69517	62964	67962	34510	12607	52255
33	28541	02029	08068	96656	17795	21484	57722	76511	27849	61738
34	11282	43632	49531	78981	81980	08530	08629	32279	29478	50228
35	42907	15137	21918	13248	39129	49559	94540	24070	88151	36782
36	47119	76651	21732	32364	58545	50277	57558	30390	18771	72703
37	11232	99884	05087	76839	65142	19994	91397	29350	83852	04905
38	64725	06719	86262	53356	57999	50193	79936	97230	52073	94467
39	77007	26962	55466	12521	48125	12280	54985	26239	76044	54398
40	18375	19310	59796	89832	59417	18553	17238	05474	33259	50595