CHAPTER 1

The Resampling Method of Solving Problems

Introduction

This chapter is a brief introduction to the resampling method of solving problems in probability and statistics. A simple illustrative problem is stated, and then the step-by-step solution with resampling is shown, using both hand methods and the computer program RESAMPLING STATS. The older conventional formulaic approach to such a problem is then discussed. The conventional analytic method requires that you understand complex formulas, and too often one selects the wrong formula. In contrast, resampling requires that you first understand the physical problem fully. Then you simulate a statistical model of the physical problem with techniques that are intuitively obvious, and you estimate the probability with repeated random sampling.

How resampling differs from the conventional approach

Recall the problem of the construction firm that owns 20 earthmoving trucks. The chance that any one truck will break down on any given day is about 1 in 10, based on past experience. You want to know the probability that on a particular day—tomorrow—*four or more* trucks will be out of action. The resampling approach produces the estimate as follows:

> We collect 10 coins, and mark one of them with a pen or pencil or tape as being the coin that represents "out-of-order;" the other nine coins stand for "in operation." This set of 10 coins is a "model" of a situation where there is a one-in-ten chance—a probability of .10 (10 percent)—of *one* particular truck being out-of-order on a given day. Next, we put the coins into a little jar or urn, draw out

one coin, and mark down whether or not that coin is the coin marked "out-of-order." That drawing of the single coin from the urn represents the chance that any one given truck among our twenty trucks (perhaps the one with the lowest license-plate number) will be out-of-order tomorrow.

Then we put the drawn coin back in the urn, shake all the coins, and again draw out a coin. We mark down whether that second-drawing coin is or is not the "out-of-order" coin, and that outcome stands for a second truck in the fleet. We do this *twenty* times to represent our twenty trucks, replacing the coin after each drawing, of course. Those twenty drawings represent one day.

At the end of the twenty draws we count how many out-of-orders we have gotten for that "day," checking whether there are *four or more* out-of-orders. If there are four or more, we write down in another column "yes"; if not, we write "no." The work we have done up to now represents one experimental trial of the model for a single day.

Then we repeat perhaps 50 or 100 times the entire experiment described above. Each of those 50 or 100 experimental trials represents a single day. When we have collected evidence for 50 or 100 experimental days, we determine the proportion of the experimental days on which four or more trucks are out of order. That proportion estimates for us the probability that four or more trucks will be out of order on a given day—the answer we seek. This procedure is an example of Monte Carlo simulation, which is the heart of the resampling method of statistical estimation.

A more direct way to answer this question would be to examine the firm's actual records for the past 100 days or 500 days to determine how many days had four or more trucks out of order. But the resampling procedure described above gives us an estimate even if we do not have such long-term information. This is realistic; it is frequently the case in the workaday world that we must make estimates on the basis of insufficient history about an event.

A quicker resampling method than the coins could be obtained with twenty ten-sided dice or spinners. Each one of the dice, marked with one of its ten sides as "out-of-order," would indicate the chance of a single truck being out of order on a given day. A single pass with the twenty dice or spinners allows us

to count whether four or more trucks turn up out of order. So in a single throw of the twenty dice we can get an experimental trial that represents a single day. And in a hundred quick throws of the twenty dice—which probably takes less than 5 minutes—we can get a fast and reasonably-accurate answer to our question. But obtaining ten-sided dice might be a nuisance.

Another fast method that needs no special equipment is using a table of "random digits," such as is found in the Appendix. If we say that the digit "zero" represents "out-of-order" and the digits "1-9" represent "in operation," then any one random digit gives us a trial observation for a single truck. To get an experimental trial for a single day we look at twenty digits and count the number of zeros. If the number of zeros is four or more, then we write "yes." We then look at one hundred or two hundred sets of twenty digits and count the proportion of sets whose twenty digits show four or more trucks being "out-of-order." Once again, that proportion estimates the probability that four or more trucks will be out-of-order on any given day.

About the *Resampling Stats* software

Thanks to the personal computer and the programming language RESAMPLING STATS, we now have a much faster way of solving problems with resampling. RESAMPLING STATS is a small set of simple, intuitive commands that get the job done quickly and efficiently, and with complete understanding on the part of the user. One may also use standard computer languages like BASIC or PASCAL to write programs that duplicate the simulation steps described above; such programs are quite cumbersome, however, and sometimes result in reliance on pre-written canned routines that are as mysterious as formulas.

The core of the program to solve the trucks problem above begins with this command to the computer:

```
GENERATE 20 1,10 a
```

This command orders the computer to randomly GENERATE twenty numbers between "1" and "10." Inasmuch as each truck has a 1 in 10 chance of being defective, we decide arbitrarily that a "1" stands for a defective truck, and the other nine numbers (from "2" to "10") stand for a non-defective truck. The command orders the computer to store the results of the ran-

dom drawing in a location in the computer's memory to which we give a name such as "A" or "BINGO."

The next key element in the core of the program is:

COUNT a =1 b

This command orders the computer to COUNT the number of "1's" among the twenty numbers that are in location A following the random drawing carried out by the GENERATE operation. The result of the COUNT will be somewhere between 0 and 20, the number of trucks that might be out-of-order on a given day. The result is then placed in another location in the computer's memory that we label B.

Now let us place the GENERATE and COUNT commands within the entire program that we use to solve this problem, which is:

REPEAT 400
Repeat the simulation 400 times

GENERATE 20 1,10 a
Generate 20 numbers, each between "1" and "10," and put them in vector a. Each number will represent a truck, and we let 1 represent a defective truck.

COUNT a =1 b
Count the number of defective trucks, and put the result in vector b.

SCORE b z
Keep track of each trial's results in z.

END
End this trial, then go back and repeat the process until all 400 trials are complete, then proceed.

COUNT z > 3 k
Determine how many trials resulted in more than 3 trucks out of order. (This can also be written COUNT z => 4 k, for 4 or more out of order.)

DIVIDE k 400 kk
Convert to a proportion.

PRINT kk
Print the result.

Note: The file "trucks" on the Resampling Stats software disk contains this set of commands.

The SCORE statement that follows the COUNT operation simply keeps track of the results of each trial, placing the number of defective trucks that occur in each trial in a location that we

usually call "z." This is done in each of the 400 trials that we make, and the result eventually is a "vector" with 400 numbers in it.

In order to make 400 repetitions of our experiment—we could have decided to make a thousand or some other number of repetitions—we put REPEAT 400 before the GENERATE, COUNT, and SCORE statements that constitute a single trial. Then we complete each repetition "loop" with END.

Since our aim is to count the number of days in which more than 3 (4 or more) defective trucks occur, we use the COUNT command to count how many times in the 400 days recorded in our SCORE vector at the end of the 400 trials more than 3 defects occurred, and we place the result in still another location "k." This gives us the total number of days where 4 or more defective trucks are seen to occur. Then we DIVIDE the number in "k" by 400, the number of trials. Thus we obtain an estimate of the chance, expressed as a probability between 0 and 1, that 4 or more trucks will be defective on a given day. And we store that result in a location that we decide to call "kk," so that it will be there when the computer receives the next command to PRINT that result on the screen.

Can you see how each of the operations that the computer carries out are analogous to the operations that you yourself executed when you solved this problem using a ten-sided spinner or a random-number table? This is exactly the procedure that we will use to solve every problem in probability and statistics that we must deal with. Either we will use a device such as coins or a random number table as an analogy for the physical process we are interested in (trucks becoming defective, in this case), or we will simulate the analogy on the computer using the RESAMPLING STATS program.

Simple as it is, the RESAMPLING STATS program called "TRUCKS" may not seem simple to you at first glance. But it is vastly simpler than the older conventional approach to such problems that has routinely been taught to students for decades.

In the standard approach the student learns to choose and solve a formula. Doing the algebra and arithmetic is quick and easy. The difficulty is in choosing the correct formula. Unless you are a professional mathematician, it may take you quite a while to arrive at the correct formula—considerable hard thinking, and perhaps some digging in textbooks. More important than the labor, however, is that you may come up with

the wrong formula, and hence obtain the wrong answer. Most students who have had a standard course in probability and statistics are quick to tell you that it is not easy to find the correct formula, even immediately after finishing a course (or several courses) on the subject. After leaving school, it is harder still to choose the right formula. Even many people who have taught statistics at the university level (including this writer) must look at a book to get the correct formula for a problem as simple as the trucks, and then we are not always sure of the right answer. This is the grave disadvantage of the standard approach.

In the past few decades, resampling and other Monte Carlo simulation methods have come to be used extensively in scientific research. But in contrast to the material in this book, simulation has mostly been used in situations so complex that mathematical methods have not yet been developed to handle them. Here are examples of such situations:

1. For a rocket aimed at the moon, calculating the correct flight route involves a great many variables, too many to solve with formulas. Hence, the Monte Carlo simulation method is used.

2. The Navy might want to know how long the average ship will have to wait for dock facilities. The time of completion varies from ship to ship, and the number of ships waiting in line for dockwork varies over time. This problem can be handled quite easily with the experimental simulation method, but formal mathematical analysis would be difficult or impossible.

3. What are the best tactics in baseball? Should one bunt? Should one put the best hitter up first, or later? By trying out various tactics with dice or random numbers, Earnshaw Cook (in his book *Percentage Baseball*), found that it is best never to bunt, and the highest-average hitter should be put up first, in contrast to usual practice. Finding this answer would not be possible with the analytic method.

4. Which search pattern will yield the best results for a ship searching for a school of fish? Trying out "models" of various search patterns with simulation can provide a fast answer.

5. What strategy in the game of Monopoly will be most likely to win? The simulation method systematically plays many games (with a computer) testing various strategies to find the best one.

But those five examples are all complex problems. This book

and its earlier editions break new ground by using this method for *simple rather than complex problems*, especially in statistics rather than pure probability, and in teaching *beginning rather than advanced* students to solve problems this way. (Here it is necessary to emphasize that the resampling method is used to *solve the problems themselves rather than as a demonstration device to teach the notions found in the standard conventional approach.* Simulation has been used in elementary courses in the past, but only to demonstrate the operation of the analytical mathematical ideas. That is very different than using the resampling approach to solve statistics problems themselves, as is done here.)

Once we get rid of the formulas and tables, we can see that statistics is a matter of *clear thinking, not fancy mathematics*. Then we can get down to the business of learning how to do that clear statistical thinking, and putting it to work for you. *The study of probability* is purely mathematics (though not necessarily formulas) and technique. But *statistics has to do with meaning*. For example, what is the meaning of data showing an association just discovered between a type of behavior and a disease? Of differences in the pay of men and women in your firm? Issues of causation, acceptability of control, design of experiments cannot be reduced to technique. This is "philosophy" in the fullest sense. Probability and statistics calculations are just one input. Resampling simulation enables us to get past issues of mathematical technique and focus on the crucial statistical elements of statistical problems.

If you intend to go on to advanced statistical work, the older standard method can be learned alongside resampling methods. Your introduction to the conventional method may thereby be made much more meaningful.

A discussion of the advantages and disadvantages of the resampling method for beginning students is given in the Appendix to Chapter 2 on "A Note to the Teacher."