previously covered notation and material. Unless there is special expertise needed, I tend to avoid using guest lecturers.

Where should a course be located and what hotel should be chosen? In the United States the best two places for short courses in biostatistics and epidemiologic methods are the Washington, D.C., area and the Atlanta area because these locations have the largest local pool of potential course attendees from government agencies in each area. Other locations of the United States have been chosen when interest in a course has been expressed by a cluster (usually at least ten) of persons from the same local source. The hotel chosen should be convenient to local course attendees and should have adequate small conference facilities.

When and how often should courses be taught? During the academic year, the only times available are during fall or spring break periods, semester breaks, and summer time. Weekend courses allow more flexibility, except that in our experience most participants prefer work-week hours. To allow for replenishing the pool of potential participants, we plan the same course in the same location no more than once every two years. Nevertheless, because we have developed a sequence of three different courses, we frequently plan the sequence in adjacent years. Within the past two years the number of invited courses has increased, so that as a consequence, we are planning fewer private business ventures in the immediate future.

When and how do we advertise? Course advertisements are usually placed about four months prior to the course date. We usually put a deadline for registration that is about two weeks prior to course date. We used to advertise in several publications, including the *American Journal of Epidemiology* and *Biometrics*; in recent years we have gotten by using only the *Epid Monitor*, which has worked for us primarily because of the reputation we have developed from past courses.

There are several other decision areas we will now very briefly mention. These include whether exercises should be given in class or for homework, what kind of teaching materials should be copied for students, what kind of AV setup should be used, and to what extent should a computer be used. Regarding exercises, we usually provide short problems with answers, but spend little class time discussing these, primarily because of the material we want to cover. We strongly recommend copying transparencies for students. This allows students to listen more carefully to lectures. We have found that using two overhead projectors is extremely valuable, especially because it allows the instructor to put a formula on one screen and an example on the other, to compare two analyses of the same data, or to allow the student to follow material being covered on consecutive transparencies. We provide examples of computer printouts and, in some cases, summary descriptions of programming command statements; however, we have found it too costly and infeasible to provide actual computers for student use. Moreover, much of the mechanics of computer use detracts from comprehension of the concepts and methods we are covering.

We consider it very important that a relaxed atmosphere be developed throughout the course; wherever possible, the use of humor and congeniality during lectures and break times is quite effective in maintaining student interest and attention.

In summary, through my experience over the past 20 years, I have found short courses to be a very effective and efficient approach to communicate fundamental, intermediate, and advanced statistical methods and concepts to a wide variety of audiences throughout the world. This experience has also been personally rewarding in terms of financial profit, travel to interesting locations, and developing new friends and professional colleagues.

[*Received September 1994. Revised November 1994.*]

Confessions of a Coin Flipper and Would-Be Instructor

Clifford KONOLD

Simulation data are used to test a student's beliefs about the relative probabilities of two sequences obtained by flipping a fair coin. The episode is used to illustrate general issues in using simulations instructionally.

KEY WORDS: Computer simulations; Intuitions; Randomness; Wait time.

Clifford Konold is Research Associate Professor, Scientific Reasoning Research Institute, University of Massachusetts, Amherst, MA 01003 and Co-Director of the Mathematics Center, TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140. The materials and research described in this article were supported by National Science Foundation Grant MDR-8954626. The opinions expressed here are the author's and not necessarily those of the Foundation. The author thanks Ruma Falk, Amy Robinson, Abigail Lipson, and two anonymous reviewers for their suggestions and helpful comments on earlier drafts.

1. INTRODUCTION

Most people spend little time flipping coins. When they do, it is not usually for the purpose of learning about coins or chance but for making a random selection. And I have never known anyone flipping a coin in such a circumstance to record how the coin landed. I have vivid memories of two occasions in my own life when a fairly important matter was decided by flipping a coin. In both cases, I can recall whether I won or lost but not whether the coin landed heads or tails. The fact that Kerrich (1961) was in prison when he dutifully recorded the results of 10,000 coin flips hints at the conditions required to motivate an empirical approach to probabilities in coin flipping.

It should come as no big surprise, then, that many people make claims about the results of coin flipping that they might discover to be incorrect if only they would conduct

TTTTTTTHHHHTHHHTHHT

Figure 1. A Sequence of Coin Flips Obtained by Repeatedly Flipping Until the Occurrence of Either HTHHT or HHHHH.

TTHTT TTHTH HTTTH HTTHT THTTH HTHHT

Figure 2. A Result of Flipping Until Either HTHHT or HHHHH Using the Block Method.

a few (thousands of) trials and keep track of what happened. Here is a question for which a bit of data might help change some minds: Suppose you were to keep flipping a coin until it landed either HTHHT or HHHHH on five consecutive flips. Which of those two sequences would you predict would occur first?

To illustrate, I repeatedly flipped a coin, letting it land on my desk, and kept going until the string of flips ended with either HHHHH or HTHHT. The results are shown in Figure 1.

If you are like many people I have talked to about this problem, you would have put your money on HTHHT, and in this instance, you would have won. Notice the four H's in the middle of the sequence. One more H there and you would have lost your bet. Many people, by the way, would be surprised to see seven T's at the beginning of the sequence. I must admit I was tempted to start over—to disregard this sequence as aberrant. Most of us think that sequences of coin flips should alternate frequently between heads and tails, more frequently, it turns out, than they typically do (Falk 1981). Indeed, that is one of the reasons people give for why they believe HTHHT is more likely in this situation than HHHHH. The sequence with all heads looks too orderly to be the result of a random process (cf. Kahneman and Tversky 1972; Konold, Pollatsek, Well, Lohmeier, and Lipson 1993).

Recently, I have been trying to teach probability by having students put their theories about outcomes of chance events to the test. Knowing that students have somewhat less determination, if not leisure time, than Kerrich (1961), I have made heavy use of the computer with which students can quickly generate and analyze data, leaving them ample time to reflect on what they observe. My hope has been that by comparing their expectations with the results of simulations, students will be motivated to reconsider their beliefs and, when necessary, replace these with beliefs in agreement with probability theory. I have been designing a probability simulation program called Prob Sim® (Konold and Miller 1994) and accompanying lab activities which make use of this basic philosophy. In this article, I describe in some detail a tutoring episode with a student and use it to illustrate a few issues I have come to regard as critical in using simulations instructionally.

2. TUTORING STUDY

In the spring of 1990, I was testing an early prototype of Prob Sim in individual tutoring sessions with undergraduates. One of these students, Kim Davis, worked as a part-time assistant in a photo lab adjacent to my office. When I asked if she would be a guinea pig in this tutorial study, she was hesitant. I knew that the following semester she would be taking a required statistics course in her major and convinced her to participate by suggesting that time spent in this study might give her a leg up. I also said I would pay her $5.00 an hour. I met with her for about 90 minutes once a week for three weeks. During these sessions, which were videotaped, we modeled several problems using the software. Here I describe only that part of our interaction that involved the flip-until problem given above.

2.1 Session One

It was near the end of our first session when I introduced Kim to the flip-until problem. I also gave her a similar problem that asked which of the two sequences was most likely if the flips were conducted in blocks, or sets, of five. An example is given in Figure 2. In this example it required six blocks for one of the target sequences, HTHHT, to occur. According to this method of flipping, the specified sequence must occur *within* a block of five flips; a sequence occurring between blocks is ignored. Notice that if strings occurring between blocks counted, I would have stopped flipping after the initial HT of the third block.

Below is a partial transcript of our discussion of these two problems, which I have cleaned up in various ways. Kim referred to the type of sequences generated in Figure 1 as "in a straight line" or "string." She firmly maintained that with the string method HTHHT was more likely than HHHHH. In the case of sampling in blocks of five, she argued that the two sequences were equally likely.

Kim: I think [HTHHT] will happen faster than [HHHHH] in a straight line.

Me: In a straight line. OK. And can you say why you expect that?

Kim: I think just from the long line, it can break into that at any point. I think it will be harder to find five of the same in a row than it will be to find more of an alternation of heads and tails.

Me: How about if I run it in groups, blocks of five?

Kim: I think that they're just as equal then.

Me: Because?

Kim: Because they're definite sequences. You can't grab from any area.

I set the computer up to simulate coin flipping under both methods. In the string method the computer mimicked the basic procedure I used to generate the sequence in Figure 1. It kept flipping until the string terminated with a specified sequence, then printed out how many flips it took before that sequence was produced. In the example given in Figure 1 it took 19 flips before HTHHT occurred. In what I will refer to as a "trial" I instructed the computer to first flip until HTHHT occurred, and then flip until HHHHH occurred. The sequence that occurred in the fewer flips was deemed the "winner" of that trial. (I conducted the trials in this two-stage fashion simply because the "draw until" command in Prob Sim allowed only one argument at a time.)

In the block method, the computer kept flipping blocks of size five until it produced the requested sequence. The

Table 1. Number of Repetitions Required to Obtain Target Sequences in Eight Trials Using String and Block Method of Sampling

	Sampling Method			
	String		Block	
Trial	HTHHT	HHHHH	HTHHT	HHHHH
1	34	104	5	24
2	37	116	34	13
3	22	29	20	15
4	48	36	22	63
5	34	50	49	63
6	26	7	19	3
7	21	101	78	109
8	85	28	59	34

program then displayed the number of blocks required to obtain that sequence. Again, a trial consisted of flipping first until HTHHT occurred and then flipping until HHHHH occurred. The winner was the sequence that occurred in the fewer number of blocks.

We conducted eight trials using each sampling method. Table 1 shows for each target sequence the required number of repetitions for each method. In the case of the string method this is the number of single flips required; for the block method, this is the number of blocks required. In the string method HTHHT occurred first in five of the trials while HHHHH occurred first in three. In the block method, the trials were split four/four between the two sequences.

I then asked Kim to evaluate her predictions in light of these results. She did not focus on the number of winners and losers, but on the number of repetitions required to obtain each sequence, saying these were "pretty close" for the blocked trials. She suggested we compute the average number of repetitions in both methods. Table 2 shows the averages she computed.

Kim: OK. That's exactly how I wanted it to turn out.

Me: How confident are you that in this case [blocks], they're equal, and in this case [string], [HTHHT] is more likely?

Kim: Very confident. I'm more confident in this [equal for block data] than this [not equal for string data].

Me: But if we had to do this [string] now one time, and you had to bet which one would occur, you'd want to bet on [HTHHT]?

Kim: Yep.

Me: So let's do it. Would you be willing to bet a dollar against my 70 cents?

Table 2. Comparison of Average Number of Repetitions to Get Target Sequences Using String and Block Sampling Methods ($n = 8$)

	Sequence	
Method	HTHHT	HHHHH
String	38.4	58.9
Block	36.3	40.5

Kim: Yep.

Me: Are you a gambling person?

Kim: I'll gamble on this.

We conducted one trial, which Kim won.

Kim: I'd keep betting like this.

Me: You'd give me that bet all day long?

Kim: Yeah, cause I'd lose some, but I think in the end I'd come out with a higher—

Me: Want to do it again, same bet?

Kim: Yeah.

Kim won again.

Me: Would you even give me better odds? Like, would you let me bet 50 cents against your buck?

Kim: Yeah.

Me: So if you were going to put down a dollar [on HTHHT] in this case, what is the least amount of money you'd let me put down [on HHHHH] before you wouldn't make the bet anymore?

Kim: If we were going to take a group of five or six, and the best average out of that won the money, I would let you go down to a cent.

I was not getting greedy here; I was trying to gauge the strength of her belief. Given her expectation that HTHHT was more likely than HHHHH, she seemed to have a good sense of the law of large numbers, that if we averaged the number of repetitions over several trials, she could be virtually certain that HTHHT would come out ahead. But I wanted to keep the experiment simple and so asked her what she would wager if we continued as above, betting on the outcome of single trials, not averages.

Kim: I think I'd go down to like 50 cents, 40 cents. Well, 50 cents.

Me: But if we take the average of eight times, like we did up here, you'd actually let me bet a penny.

Kim: Yep.

Kim was to return for two more sessions. I planned to continue betting on trials to see how long it would take her to abandon the belief that HTHHT was more likely than HHHHH using the string method. The combination of being wrong and losing money would certainly force her to change her prediction and provide the motivation for theory revision. That was precisely the process I was interested in investigating.

Before describing what happened in those sessions, I need to make an embarrassing admission that may help to calm the rage some of you are experiencing by now. I was wrong about this problem, and Kim was right: HTHHT *is* more likely than HHHHH to occur first if run in a string, but equally likely if done in blocks of five. In fact, because the odds she gave were not far from fair odds, she spared me from losing big. Nevertheless, the tables had been turned, and I was the unwitting subject of my own research. The question was: How much data would we need to collect before I changed *my* mind? The answer would be: A lot.

I am sure the results to this point had little impact on my belief. Anyone who has used probability simulations instructionally is quite used to getting "bad" results with

small samples. Consequently, I gave little attention to the difference of 20 between the two averages in the string method in Table 2. Indeed if, as I thought, the two were equally likely, a difference at least that large in either direction would occur by chance about a quarter of the time.

2.2 Session Two

The transcript of our second session picks up with me setting up the computer to continue the same gamble. I questioned her as I entered the information into the program.

Me: I don't remember your sequence.
Kim: Mine was HHTHT.
Me: I'll let you play any number [sequence]. And the bet is, I put up 60 cents and you put up a dollar, and we do 10 times. And we want a block size of 5. Right?
Kim: But we didn't chunk it out. It was just [pulls hands apart to suggest a long line].
Me: Oh, sorry, it was in the line. That's right.
Kim: I'll lose *your* way!
Me: And we want to run till HHTHT.
Kim: I actually think it was HTHHT.
Me: You want to change it?
Kim: Yeah, I do.

Given my misunderstanding of the situation, it did not matter which sequence she bet on, or whether we conducted trials in a string or in blocks. This could explain my poor recall of the specific sequence she had bet on and of the sampling method we had used. Under her theory both the specific sequence and the sampling method make a difference. I am not sure how to maintain my innocence in remembering having to pay Kim 60 rather than 70 cents when I lost, a slip she either missed or chose to ignore.

We ran ten trials. Overall, her sequence came up first six times. Had we used the original betting odds, I would have lost 20 cents. As it was, I was up 40 cents.

Kim: It was close that time, but this [HTHHT] came out best. I still strongly believe in my theory.

My confidence was unshaken as well. I had just grown richer, and the six/four split was perfectly within reason. Time, I thought, was on my side. We ran ten more trials, with the same win/loss results.

Kim: So, once again, I only beat you by one [two], but I owe you 80 cents now.
Me: Want to do it again [anxious for the redemptive fire of the law of large numbers]?
Kim: [Hesitates]
Me: Do you want to change your odds from—
Kim: Let's do it again. I want to kick your butt once [anxious for the redemptive fire of the law of large numbers].

At this point she still appeared to believe not only in her theory, but also that the odds she had given were to her advantage. We ran another ten trials, this time each of us winning five.

Me: So, I win 2 dollars on that round. That's $2.80 [total].
Kim: Yeah.
Me: "Yeah" what?
Kim: Stop.

2.3 Session Three

As our second session ended, Kim's confidence was on the ropes, and I expected in our final session to observe what I was most interested in, how she formulated a new understanding of the situation. But by the next week her confidence had returned—she was ready to repeat the bet. Unfortunately, we waited to the end of our session to continue the betting, and the tape ran out after we had done only one set of ten trials. Her sequence came out on top seven times, netting her $1.20. Of course, I figured that this was just chance being uncooperative again. The last thing on the tape is me instructing her that the two sequences are "in fact" equally likely. I apparently wanted to see how she would accommodate this information, and was doubtful that in the time remaining we could collect enough data to erode her now-growing confidence. In spite of my saying this she said she would continue the bet, and we ran three more sets of ten for which I have only written records. Her sequence won a total of 20 times to my 10. Now, as I forked over the $3.20 I had lost that day, it was *my* confidence on the ropes. I had been playing for quite a while with what I thought were fantastic odds, and yet had lost money.

3. THE AFTERMATH

After this last session I returned to my office curious enough to conduct an additional 100 trials. The sequence HTHHT came up first on 61 of the occasions. I did not compute it at the time, but a difference this extreme would occur only about 2% of the time if the two sequences were equally likely.

It was while running the 100 repetitions that I remembered a problem with which Warren Page had stumped me a few years previously: Which would be the most likely result, HH or HT, if you kept flipping a coin until you got one or the other? I remember at first being surprised on discovering that HT was more likely, but it was not hard to see why. With HH, every time you get a T, you are back to square one: You need to flip 2 H's. But with HT, as soon as you get one H, you are "locked in": A T on the next flip will bring success. If instead you get an H, you are still only one T away from success. With this problem in mind I sketched the diagrams in Figure 3, which convinced me that Kim's intuition was correct. Even though there is no stage in the generation of HTHHT in which you can become locked in, it is clear that it is harder with that sequence than it is with HHHHH to get sent, Shoots-and-Ladders style, back to the beginning.

I ran additional simulations to estimate the average number of flips required to obtain each sequence. In 400 trials the average for HHHHH was 62.16 flips compared to 34.6 flips for HTHHT. My requests for assistance in finding a formal solution were answered by Ruma Falk and Rolf Biehler. They both directed me to the work of Engel (1975) who developed a method of computing these average "wait-times" from just the kind of directed graphs

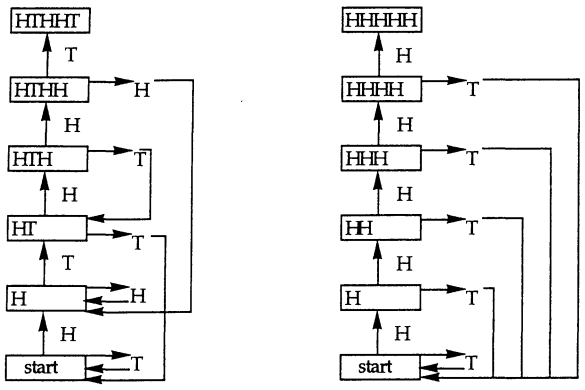

Figure 3. Directed Graphs for Flipping HTHHT and HHHHH Using the String Method.

I had constructed. Using his method the theoretical wait-time values can be determined to be 62 for HHHHH and 36 for HTHHT [see also Hoffman (1978)]. I shared these findings with a deservedly proud undergraduate tutor of mine, Kim Davis.

4. DISCUSSION

I am still convinced that simulations offer a fruitful approach to teaching probability, but designing effective instruction around them is trickier than I had imagined. In the remainder of this article, I make a number of recommendations about using computer simulations in teaching probability. I use the above episode as an example to help stress the point that various difficulties associated with reasoning from simulation data are not peculiar to students, but are obstacles we all face in trying to link theories with data.

4.1 Data Are Not Forceful

The results of a simulation do not come knocking at one's door announcing, "This is the data. Open up." Speaking of his 10,000 coin flips and theorems regarding random walks later elucidated by Feller, Kerrich (1961, p. 19) tells of being suddenly struck with the realization that he had "collected data that contained these startling results twenty years ago and had never so much as glimpsed them." The histories of science and mathematics are filled with examples where committed practitioners not only failed to see what seemed obvious to their successors, but also, some would argue, "distorted" or "explained away" what they saw. Gilbert and Mulkay (1984) describe a relatively recent debate in biochemistry in which rival researchers accused one another of bad faith for ignoring just the data that would lead them from darkness to light. One does not need to hold the radical view that data are "theory laden" (i.e., dependent on theory for their very existence) to accept that forming or testing a theory on the basis of data is, by nature, always problematic. And this is the plight of data within the domain of science which places a premium on them; data have even less chance of altering beliefs outside the practice of science, where they are casually collected, seldom recorded, and selectively attended to. Prior to my sessions with Kim, I had conducted many trials similar to the ones we conducted together without ever noticing anything unusual. I was virtually certain of what I would observe, so I never bothered looking carefully or even recording the results. Also, in spite of the fact that in the sessions with Kim I had to examine the length of the strings to determine the winner of each trial, I did not notice the large discrepancy in the averages of the two wait times, both because of my expectations and also because of the variability among trials which hides the difference from the casual observer.

The other extreme is having no expectations of what one will observe in data. This mind set also is not conducive to learning from data because there is no experience of surprise that can serve to focus attention. To maximize the possibility that students attend to data, I have them make predictions about what they expect to observe before collecting data, and ask them to be as explicit as possible about the reasons underlying their predictions. Having agreed with Kim to pit our theories against one another, I was finally in a position, as was she, to notice discrepancies between expectations and actual experience. Even then, however, the data forced neither of us to concede at any point during our three sessions. But having recorded and compared trial results to our predictions, we were poised to feel discomfort as we considered the possibility that we might be mistaken.

4.2 Attention is a Limited Resource

In my early attempts in designing simulation activities, I had students make predictions about a number of events, and keep track of each of these in each sample they drew. My intention was to maximize what they learned from each sample of data. But many students would either become overwhelmed or lose interest, and so I abandoned this practice and now have them focus on only one question at a time. In the interview with Kim the gambling may actually have diverted attention from the major question. I was most interested in having her decide whether or not the two sequences were equally likely, not in having her decide if the odds she had given were fair. Note that at the end of session two her sequence had outperformed mine, and yet she appeared ready to abandon her belief perhaps because she was losing money. Setting even odds may have eliminated this distraction.

4.3 Rarely are Enough Data Collected

Several years ago, Cohen (1962) analyzed 70 studies published in a reputable psychological journal and concluded that researchers were using sample sizes that were so small that they had only a 50% chance of rejecting a false null hypothesis. One of the anonymous reviewers of this article called my attention to the study of Freiman, Chalmers, Smith, and Kuebler (1978) who looked at clinical studies designed to determine the effectiveness of new medical therapies. They examined 71 studies that reported no therapeutic effect and found that, of these, 50 had greater than a 10% chance of labeling as ineffective a therapy that, in reality, resulted in a 50% improvement.

I have not formally surveyed probability curricula using simulation, but my guess is that frequently there are not enough data collected to warrant drawing conclusions. Before the availability of the computer, when classroom

data were collected by flipping coins or drawing marbles from containers, this was an often unavoidable problem. There is a tendency in using the computer, however, to easily underestimate the time required to adequately simulate a particular problem. Depending on the problem under investigation, the speed of the computer, and the design of the simulation software, computer simulations may still be too slow to permit drawing sufficient data in the allotted time. My guess is that with increases in computing speed, we will graduate to more complex problems, and thus still frequently be drawing too little data.

In retrospect, I clearly did a poor job preparing for my interviews with Kim—given the time I had set aside, there was only a slim chance that results we collected would provide a compelling basis for changing either of our minds. Assuming I was correct in my belief that there was no difference in the wait times for flipping HTHHT versus HHHHH, what was the chance of collecting sufficient data to "decide" the issue? In the second session with Kim we conducted 30 trials. The time to set up, run, and record these trials was about 10 minutes. We were not explicitly pooling data across sessions, so I must have been assuming that these 30 trials would supply adequate information to pose a serious challenge to Kim's theory. What results would have led Kim to question her belief? We never established decision points, but note that HTHHT won 17 out of 30, and Kim did not seem too encouraged by this result. So, let us assume the following symmetric partition of the number of times HTHHT appears first in the 30 trials:

18–30: support for Kim's theory
13–17: support for my theory
0–12: support for some other theory.

What was the probability of each of these three partitions given my assumption of equal probability? According to the binomial distribution, with $p = .50$, the chance of getting data consistent with my theory is almost 64%, which leaves roughly 18% chance of obtaining data consistent with Kim's theory and 18% chance of befuddling both of us.

Had I thought this through before the session with Kim, I would have either planned on conducting many more trials than we did or changed the random variable we investigated. Had we compared the average wait-times for the two sequences, as Kim had initially suggested, we could have had my mind changed in 5 or 10 minutes of sampling. The important point is that when designing classroom simulations, you need to consider carefully the sample size (and thus time) required to permit arriving at a reasonable conclusion. And whatever you determine this sample size to be, triple it.

4.4 Variability is Typically Ignored

Simulations are frequently used, as they were here, to determine the relative probabilities of two or more mutually exclusive events. For example, suppose you had students predict which of the two sequences HTHHT versus HHHHH is more likely in five flips of a coin (i.e., using the block method). Many will predict that HTHHT is the more likely. It seems reasonable that they could learn otherwise from conducting simulations. Imagine that a student runs 10,000 repetitions of this experiment in which there were 312 occurrences of HHHHH and 320 of HTHHT. What is this student to conclude from this? These frequencies are probably closer than the student would have predicted before collecting data. But those who believe that HTHHT is more likely might argue that these data support that belief. This is not a case of drawing a sample that is too small, but of the students not having available a basis for evaluating the magnitude of an observed difference. This problem is sometimes finessed by having students compare relative frequencies, in this case .0312 versus .0320. Students may be more likely to agree that the latter values are nearly equal, but they are still evaluating differences without regard to the variability in the sampled values.

Table 3. Frequency of Occurrence of HHHHH Versus HTHHT in Ten Repetitions of 1,000 Trials in Which the Coin Was Flipped Five Times in Each Trial

	Rep. No.									
Sequence	1	2	3	4	5	6	7	8	9	10
HHHHH	33	28	24	34	34	28	26	36	28	29
HTHHT	20	36	26	26	24	37	35	36	32	29

Although I do not introduce formal measures of variability in introductory courses, I do have students conduct multiple repetitions of a trial of some sample size. The different outcomes of each repetition reveal the variability inherent in the sampling process and give some sense of the magnitude of that variability for the given sample size. In the example here, instead of having students conduct 10,000 trials in one step, I would have them draw 10 repetitions of samples of size 1,000. If these are recorded, as in Table 3, many students will pay little attention to the variability.

The variability over repetitions is more salient when the results are plotted, as in the histograms in Figure 4.

Having students compare (or pool) their results with those from others in the class calls further attention to the variability in results. If instead of frequencies they

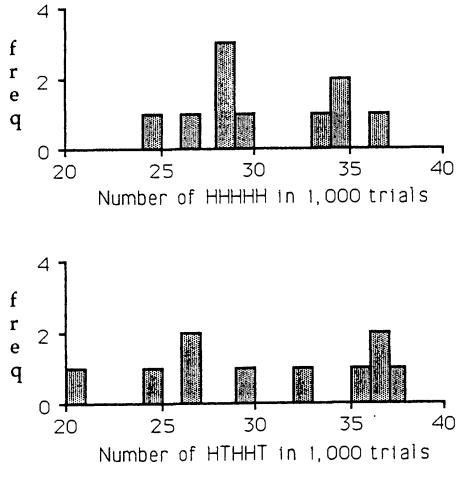

Figure 4. Histogram Display of Results Shown in Table 3.

plot relative frequencies, students can compare variability for repetitions with different sample sizes and develop a general sense of the law of large numbers.

4.5 The Focus Should be Sense Making

To discover through simulations using the string method that HTHHT was more likely than HHHHH to appear first was not the end, but the beginning, of my inquiry. Had we found them instead to be roughly equal, I doubt Kim would have blithely accepted that. She would have puzzled over why that was, and I would have encouraged her to pursue the question until it made sense to her. In short, simulations should not be used to replace the traditional theory-driven approach to teaching probability with a purely empirical approach. Some students are only too happy to draw sweeping conclusions from their simulations. I was stunned to read one student's summary of what he had learned that week in a class I was teaching: "I used to think that with coin flipping heads and tails were equally likely. But through using the computer I've learned that a tails is more likely."

Simulations give us a way of testing *our* theories, and these theories should remain the primary focus. I stress the word "our" because it is fairly common to use simulations to validate theoretical probabilities or probability postulates. My own belief is that this approach has a good chance of leaving untouched the informal notions students bring into the classroom. The approach I have taken is to encourage students to articulate their informal theories, to make predictions from them, and to use the results of simulation to motivate the need for alternative explanations (Konold 1994). As a result, I devote considerable classroom time to discussing students' expectations and theories before having them collect data. After they have collected data and reconsidered their theories, I have them as a class discuss their findings. I make sure that an acceptable theoretical explanation is one of the options under consideration in this final discussion. The ratio of classroom time spent at the computer to that spent in discussion has changed from about 2 to 1 in my early attempts at using simulation to 1 to 2 now. It is during these classroom discussions, and not usually at the computer, where understanding finally develops.

Let me quickly add that what I regard as understanding is not synonymous with having a formal solution to some problem. I do not mean to minimize the importance of formalization. I was not satisfied that I had solved the flip-until problem until I learned from Engel (1975) how the recursive structure of the directed graphs in Figure 3 could be mastered and obtained theoretical validation of the simulation results. However, understanding how to apply the formula is not what enlightened me; it was thinking about the directed graphs themselves. Examining these, I could see why a string of random outcomes was more likely to be closer to the final state HTHHT than to the final state HHHHH.

Similarly, I cannot remember how I came to understand that all coin-flipping sequences of equal length are equally likely, but I am confident it was not by thinking about the implications of $(1/2)^n$. More likely, the understanding developed over time as a result of looking at tree graphs, of thinking about the implications of believing that the probability of each outcome on each flip remained 1/2, and of noticing that my expectation that a mixed-up sequence was more likely than a patterned one was correct if I considered unordered strings (e.g., three H's, two T's) rather than ordered sequences (e.g., HTHHT). Of course, this is precisely the understanding that I overgeneralized when considering the string version of the problem. But it is this kind of understanding that permits us to generalize at all with some degree of success.

Moreover, understanding does not typically arrive suddenly like a newborn and set up permanent residence. More like a teenager, it pops in and out. After I thought I had come to terms with the flip-until problem, the following dilemma set me back momentarily. Suppose I flipped a coin 1,000 times and wrote down the results in one long string. I could search for occurrences of HHHHH and HTHHT by sliding a "window" along the string that allowed me to see only five characters at a time. If I started at the beginning of the string and advanced the window one character at a time, I could view 996 events of length 5. I am convinced that in this sample the expected number of occurrences of HHHHH, HTHHT, or any other sequence of length 5 is $996\ (1/2)^5$. How can this be reconciled with the fact that, sliding the window along, I expect to encounter the first instance of HTHHT before encountering HHHHH?

[*Received June 1993. Revised November 1994.*]

REFERENCES

Cohen, J. (1962), "The Statistical Power of Abnormal-Social Psychological Research: A Review," *Journal of Abnormal & Social Psychology*, 65, 145–153.

Engel, A. (1975), "The Probability Abacus," *Educational Studies in Mathematics*, 6, 1–22.

Falk, R. (1981), "The Perception of Randomness," in *Fifth Conference of the International Group for the Psychology of Mathematics Education*, Grenoble, France, pp. 222–229.

Freiman, J. A., Chalmers, T. C., Smith, Jr., H., and Kuebler, R. R. (1978), "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial," *The New England Journal of Medicine*, 299, 690–694.

Gilbert, G. N., and Mulkay, M. (1984), *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse*, Cambridge, U.K.: Cambridge University Press.

Hoffman, N. (1978), "Some New Ways of Solving a Coin Tossing Problem," *The Two-Year College Mathematics Journal*, 9, 6–10.

Kahneman, D., and Tversky, A. (1972), "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 3, 430–453.

Kerrich, J. E. (1961), "Random Remarks," *The American Statistician*, 15, 16–20.

Konold, C. (1994), "Teaching Probability Through Modeling Real Problems," *Mathematics Teacher*, 87, 232–235.

Konold, C., and Miller, C. (1994), "Prob Sim®: A Probability Simulation Program," Santa Barbara, CA: Intellimation Library for the Macintosh.

Konold, C., Pollatsek, A., Well, A. D., Lohmeier, J., and Lipson, A. (1993), "Inconsistencies in Students' Reasoning about Probability," *Journal for Research in Mathematics Education*, 24, 392–414.