UNDERSTANDING THE LAW OF LARGE NUMBERS

Clifford Konold, Arnold Well, Jill Lohmeier, and Alexander Pollatsek
University of Massachusetts, Amherst

Twelve students answered questions involving the distribution of sample means both before and after an instructional intervention. Correct performance improved on these problems but dropped on problems having to do with the distribution of samples.

Perhaps the most fundamental concept in inferential statistics is the sampling distribution. The most basic property of sampling distributions is the law of large numbers (LLN) — that statistics of larger samples (e.g., means) more closely approximate the corresponding parameters of populations and are thus less variable than those of smaller samples. Mastery of this law can carry a student a long way in statistics. Early research by Kahneman and Tversky (1972) suggested that people had fairly poor intuitions about the LLN. Below is a problem we adapted from them.

> Post-Office Problem (Tails version). When they turn 18, American males must register at a local post office. In addition to other information, the height of each male is obtained. The national average height for 18-year-old males is 69 inches (5 ft. 9 in.). Every day for one year, 5 men registered at a small post office and 50 men registered at a big post office.
>
> At the end of each day, a clerk at each post office computed and recorded the average height of the men who registered there that day.
>
> I would predict that the number of days on which the average height was more than 71 inches (5 ft. 11 in.) would be
>
> a. greater for the small post office than for the big post office.
> b. greater for the big post office than for the small post office.
> c. There is no basis for predicting for which one, if either, it would be greater.

Only 15-20% of the undergraduates in a study by Well, Pollatsek, and Boyce (1990) answered this problem correctly (option *a*). Roughly 65% answered *c*, suggesting that they believed sample size has no effect on the variability of the mean. Kahneman and Tversky (1972) concluded that this answer is indicative of the "representativeness heuristic," according to which the likelihood of drawing a particular sample is judged by considering how similar the sample is to the parent population. In this problem the two samples differ only in size, but because sample size is not a feature of a population, it is not used as a criterion to judge the similarity between the samples and the population. Accordingly, people tend to think even small samples will adequately represent the critical features of the population.

We have been designing software that demonstrates various features of the LLN and testing it in tutoring interviews with undergraduates. The software permits drawing samples of various sizes from a population of elements and plotting the means of those samples in a

sampling distribution. By comparing the empirically constructed distributions of different sample sizes, one can observe that the means of larger samples tend to cluster more closely around the population mean than do means of smaller samples. In an earlier study we conducted, students received instruction on sampling distributions of the mean. Computer simulations were used to construct sampling distributions based on different sized samples. This intervention increased correct responses to questions concerning the variability of means as a function of sample size by a factor of four. However, when we asked questions about the distribution of elements within a sample, many subjects over generalized what they learned about statistics to the distribution of the samples themselves (Lohmeier, 1992; Well, Pollatsek, & Boyce, 1990). In this study, we look more closely at the nature of students' responses to various problems before and after similar computer-enriched instruction.

Method

Tutoring interviews were conducted with 12 undergraduates recruited from psychology courses at the University of Massachusetts. The individual sessions lasted about 75 minutes and were videotaped. Before introducing the simulation software, students solved two problems. The first problem "Post-office tail" was identical to the one given above. The second problem, "Post-office lure," is given in abbreviated form below:

> Post-office lure. When they turn 18 . . . 50 men registered at a big post office [same paragraph as in the Post-office tail problem].
>
> At the end of one day, a clerk at each post office computed the percentage of men who registered that day who were less than 66 inches tall.
>
> I would expect that the percentage of men less than 66 inches would be [same three options as in the Post-office tail problem].

This problem was constructed to test the degree of understanding of the LLN after instruction. Note that the lure problem does not ask about the distribution of sample *means* from the two post offices over the period of a *year*, but about the distribution of the *individual men's heights* at the post offices on a particular *day*. Because the expected value for the requested percentage in both samples equals the corresponding percentage in the population, the correct answer to the lure problem is *c*.

For each problem, students were instructed to first read the problem aloud and then to solve it while "thinking aloud." The interviewer (Lohmeier) encouraged students to vocalize their thoughts and also probed them when responses were unclear. After selecting and justifying an answer, the page on which the problem was written was turned over, at which time the student was asked to "restate the problem in your own words."

Following the instructional intervention, which lasted about 40 minutes, students again solved the Post-office tail and Post-office lure problems under the same instructions. These

were followed by three transfer problems. The first of these, "Post-office center," was similar to the Post-office tail problem. The only difference was that the former asked about "the number of days on which the average height was within 2 inches of the national average (69 plus or minus 2 inches)" rather than about the frequency of means in the tail of the sampling distribution. Thus, the correct answer in this case is *b*, that the larger post office will likely record more such days. The other two transfer problems used different cover stories but were structurally compatible with the Post-office tail problem ("Geology tail") and Post-office lure problem ("Treasury lure").

Results

We discuss three aspects of students responses: a) their answers, b) rationales for their answers, and c) the accuracy of their restatements of the problem.

Answers to Problems

Answers (along with students' restatements of the problem) are summarized in Tables 1 and 2. Table 1 shows performance on the tail and center problems, Table 2 on the lure problems.

Table 1. Summary of performance on tail and center problems

Before				After											
Post Office Tail				Post Office Tail				Geology Tail				Post Office Center			
S#	ans	qty	object	S#	ans	qty	object	S#	ans	qty	object	S#	ans	qty	object
8	S	#	mean	8	S	#	mean	9	S	#	mean	7	L	#	mean
1	E	#	mean	1	S	#	mean	1	S	%	mean	8	L	#	mean
11	E	%	mean	7	S	#	mean	2	S	%	mean	10	L	P	mean
3	E	?	mean	2	S	P	mean	7	S	%	mean	3	L	?	mean
4	E	?	mean	10	S	P	mean	4	S	P	mean	11	L	%	elem
5	E	?	mean	9	S	%	mean	12	S	?	mean	9	L	%	?
9	E	?	?	6	S	?	mean	8	S	?	?	5	L	%	?
7	E	#	elem	12	S	#	?	6	S	%	elem	6	L	P	?
2	L	#	elem	3	S	?	?	11	E	%	mean	1	L	•*	•
12	L	#	elem	11	E	#	mean	5	E	P	mean	4	L	•	•
6	L	P	elem	5	L	#	mean	10	E	?	mean	2	S	#	mean
10	L	P	?	4	L	#	elem	3	E	%	?	12	S	?	mean

* Missing values indicate that the student was not asked to restate the problem.

We look first at answers, coded under columns labeled "ans." On the tail problems, the correct response is that S (the smaller sample) is more likely; for the center problem, L (the larger sample) is correct. Correct answers are grouped at the top of a column and enclosed in a

box. The number of correct answers on the Post-office tail problem increased from 1 to 9 over instruction, and remained reasonably high (8 and 10) on the 2 transfer problems. Moreover, for the most part, students who responded correctly had not simply learned to answer "smaller," since they gave the correct answer "larger" when the question concerned the frequency of values in the center as opposed to the tails of the sampling distribution. Judging from these results alone, one would conclude that the instruction had been quite successful.

Table 2. Summary of performance on lure problems

Before				After							
Post Office Lure				Post Office Lure				Treasury Lure			
S#	ans	qty	object	S#	ans	qty	object	S#	ans	qty	object
1	E	%	elem	3	E	%	elem	9	E	%	elem
3	E	%	elem	9	E	%	elem	1	S	%	elem
4	E	%	elem	11	E	%	elem	2	S	%	elem
5	E	%	elem	12	E	#	elem	6	S	%	elem
11	E	%	elem	5	S	%	elem	8	S	%	elem
9	E	%	?	8	S	%	elem	10	S	%	elem
7	E	•	•	1	S	#	mean	12	S	?	elem
6	L	%	elem	2	S	?	mean	4	S	?	mean
8	L	%	elem	10	S	?	mean	3	L	%	elem
10	L	%	elem	4	L	%	elem	5	L	%	elem
12	L	%	elem	6	L	%	elem	7	L	%	elem
2	L	%	?	7	L	%	elem	11	L	#	elem

Performance on the lure problems appears to show a negative effect over instruction. Before instruction, 7 students correctly answered the Post-office lure; after instruction, only 4 correctly answered the same problem, and only 1 correctly answered the Treasury lure. These results replicated findings of an earlier study (Well, Pollatsek, & Boyce, in preparation), suggesting that after instruction students have some understanding of the LLN, but are not yet able to fully discriminate between situations when it is and is not applicable.

Restatements of Problem

After students had given an answer, they were asked to restate the problem in their own words. These statements gave an indication of how students encoded the problem, providing a context for further evaluating their answers. We were interested, in particular, in the degree to which incorrect answers resulted from encoding errors. If, for example, students misinterpreted a tails problem as asking for the distribution of elements in a sample (i.e., as a lure question), the answer E (equal) would be correct given their interpretation. Entries in Tables 1 and 2 under the columns headed "qty" (quantity) and "object" were obtained by

coding students' restatements. The object column shows whether the student said the problem was asking about a) means of the two samples or b) sample elements (elem). The qty column codes a student's restatement according to the quantity sought: number (#), percent (%), or probability (P). For example, below is the restatement by S6 of the Post-office tail problem before instruction, which was coded as "P, elem" based on the underlined phrases:

> What the odds would be that someone 5'11" would be, would it be greater that they'd go to the smaller post office — well, that you'd get someone that height at the smaller post office or the larger post office ....

Overall, the accuracy of encoding the tail problems improved over instruction. Whereas before instruction only 3 students correctly restated the tail problem, after instruction 8 gave correct restatements on the Post-office tail problem and 7 on the Geology problem. (Though "number" was the specified quantity for the tails and center problems, "percent" and "probability" were also considered correct encodings.) Moreover, although before instruction there were 5 clearly incorrect (as opposed to incomplete) responses on the Post-office tail problem, in each of the post-instruction problems only 1 student's restatement was incorrect.

Nine of the restatements of the lure problem were correct before instruction, and none were incorrect. ("Number" was considered an incorrect encoding of the quantity on lure problems.) While the number of correct restatements remained about the same after instruction (8 and 9), the number of incorrect responses grew to 4 and 2, with half of these reinterpretations involving means rather than elements. This is further indication that performance on questions involving distributions of the mean improved, to some extent, at the expense of performance on questions involving distributions of samples.

Our major interest in coding restatements was to determine to what extent incorrect answers resulted from related encoding errors. People may respond "equal" (E) to the problems about distributions of means because they interpret them as questions about the percentage of elements in the sample distribution above or below some value. We obtained little evidence for this hypothesis. Of the 12 E answers given over the 4 sampling-distribution problems, in only 1 case was the problem restated in terms of elements.

Students who responded L to the tail problems may also have been interpreting the problems as asking about sample elements. This answer would be reasonable if they misinterpreted the question as asking not only about sample elements (rather than means), but also about the number (rather than percent) of elements. The responses of S2 and S12 on the pretest were consistent with this interpretation. However, they gave the same answer to the lure problem before instruction, and yet correctly interpreted the questions as concerning percents. We say more about this below.

Rationales

We can make only a few comments about rationales students gave. First, after instruction students who correctly answered the sampling-distributions problems gave rationales that were, for the most part, consistent with the LLN. Summing over the 3 after-instruction problems, 22 of the 27 correct answers (81%) were accompanied by rationales that were consistent with the LLN, as exemplified by the rationale of S9 to the Post-office tail problem after instruction.

> I don't think there is any basis again . . . no, no, no. These are means. It would be greater for the small post office. This is a shift that I didn't make from talking about a population distribution to the distribution of sample means. This is the same question I read before, isn't it? And they're talking about two different distributions. I did not make that shift . . . . When the sample is smaller in a distribution of sample means, the standard deviation will be greater from the mean, and for any given number away from the mean, there will be a greater number of points.

This excerpt suggests this student learned a new distinction as a result of instruction. This is the only student who, after instruction, answer all problems correctly.

Students who incorrectly answered E to the sampling-distribution problems for the most part argued as did S3: "I just don't think there is enough information for me to answer that." This same argument was also given by most students who correctly answered E to the lure problems. The vagueness of this argument suggests that most correct responses to the lure problem before instruction were not based on the understanding that, on average, the two samples would have the same percentage of elements in the specified area. It also suggests that incorrect responses of E on the tail problems were not based so much on the representativeness heuristic, but on the belief that there is not enough information given to make a prediction — a variety, perhaps, of an equal-ignorance argument.

Finally, the majority of students (78%) who gave incorrect answers of L to the lure and tail problems justified these with what we refer to as the "more-is-more" argument. This is the belief that the larger sample will have larger x, where x can be scores, means, percentages, variability, etc. For example, below is the justification given by S2 to the Post-office tail problem before instruction:

> There would be more people going to the bigger post office. So the average number of days would be greater for the big post office . . . . If they have more people going to the big post office, then the average is going to come out higher than the small one.

For a few students, this belief may consist primarily of a failure to understand the difference between number and percent. Some made no apparent distinction between the two; others distinguished number and percent but vacillated between them during the interview;

still others apparently recognized some difference but, reasoning from the more-is-more notion, thought it unimportant, as illustrated, again, by S2:

> I guess it would be greater for the big post office. When you're talking about percentage, you're talking about a number of people .... You'd have less people going to the small post office and more at the big post office ... more of a chance at the big post office.

The more-is-more rationale may explain L answers on lure problems by students who nevertheless correctly encoded the problems as dealing with percents rather than numbers.

Conclusions

We should emphasize that our objective in these tutoring studies is not to design a one-hour intervention for teaching the LLN. It may even be overly ambitious to expect much understanding after a semester of statistics instruction. Rather, we are interested in exploring related intuitions that exist prior to instruction and the nature of difficulties that arise as various concepts are introduced as part of instruction. Our research suggests that students' difficulties learning the LLN are not based on incompatible intuitions, such as the representativeness heuristic, and that, indeed, by using the computer to demonstrate the construction of sampling distributions, most students quickly develop some understanding of why means of larger samples are less variable than those of smaller samples. However, poor understanding of percentages, and whatever other misunderstandings support the more-is-more rationale, present a barrier for many students. Finally, we used lure problems in this study to test the degree of understanding of the LLN. We are not satisfied that students understand the LLN as long as they continue applying it to distributions of sample elements. While it was not part of instruction in this study, we would expect that it would facilitate understanding if during instruction students were given problems which deal both with distributions of means and with distributions of samples, addressing explicitly why the LLN applies in one case but not in the other.

References

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430-453.

Lohmeier, J. H. (1992). The effects of training of understanding the law of large numbers. Unpublished MS thesis, Department of Psychology, University of Massachusetts, Amherst.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Journal of Organizational Behavior and Human Decision Processes*, 47, 289-312.

Well, A. D., Pollatsek, A., & Boyce, S. J. (in preparation).