Learning About Sampling : Trouble at the Core of Statistics

Andee Rubin, Bertram Bruce and Yvette Tenney - Cambridge, Massachusetts, USA

1. Overview

For the past several years we have been studying the teaching and learning of statistical reasoning, in the context of a United States National Science Foundation-supported project called ELASTIC. One of the most impressive realisations we have had is just how difficult the basic concepts of sampling and statistical inference can be for students - and often teachers - to grasp. This paper explores some of the underlying conceptions and heuristics students bring to the study of statistics, and makes some initial hypotheses as to how these approaches might complicate students' learning the foundations of statistical inference.

We have organised our investigations around a set of concepts about sampling that are central to understanding statistical inference. One way of stating the central idea of statistical inference is that a sample gives us *some* information about a population - not nothing, not everything, but something. In practice, this allows us to put bounds on the value of a characteristic of the population - usually either a proportion or a measure of centre (mean or median), but not to know precisely what that characteristic is.

This kind of reasoning follows from the somewhat antithetical notions of sampling representativeness and sample variability. *Sample representativeness* is the idea that a sample taken from a population will often have characteristics similar to those of its parent population. Thus, the proportion of girls in a classroom is *likely* to be close to the proportion of girls in the entire school, and the mean family size of the students in the class is *likely* to be close to that of the whole school. *Sample variability* is the contrasting idea that samples from a single population are not all the same and thus do not all match the population. Thus, some classrooms in a school are *likely* to have many more girls than boys, even if the school population is evenly divided.

One of the keys to mastering statistical inference is balancing these two ideas, interpreting more precisely the meaning of "likely" in each. Because they are contradictory when seen in a deterministic framework, it is possible that people have a tendency to over-respond to one or the other in various contexts. Over-reliance on sample representativeness is likely to lead to the notion that a sample tells us *everything* about a population; over-reliance on sample variability implies that a sample tells us *nothing*.

In order to investigate students' naive conceptions of sampling representativeness and variability, we interviewed twelve senior high school students who had never taken a statistics course. The interview consisted of six open-ended questions related to sampling and statistical inference and took approximately half an hour. The data consist of transcribed audiotapes of the interviews plus the scratch paper students used to work on the problems. The complete texts of the questions are available from the authors.

Our analysis of their responses indicates that students have inconsistent models of the relationship between samples and populations, even in problems where the underlying mathematical models are all binomial distributions. In some situations, the notions of sample representativeness hold sway, in others, those of sample variability are more powerful. Sample size does not seem to operate appropriately to separate the two; in fact, in the problems we analyse here, sample representativeness appears to be a stronger guiding factor in the problem with the smaller sample size. In the remainder of the paper, we analyse two of the problems students encountered from this single perspective of sample representativeness and variability and note the inconsistency of their responses.

2. Gummy bears question : effects of focussing on sample representativeness

Students were told that the Easter Bunny made many packets of six Gummy Bears from a large vat containing two million green and one million red Gummy Bears and distributed them at an Easter Parade. Students were first asked how many green Gummy Bears they thought might be in their own packet. They then estimated how many kids out of 100 would have that same number of green Gummy Bears. Finally, we asked them to specify the entire distribution by answering the questions, "How many kids out of 100 had N green Gummy Bears in their packet?" for N = 0 through 6.

The theoretical model for this situation is a binomial distribution with p = .66 and sample size 6. The distribution peaks at the sample containing 4 green and 2 red Gummy Bears, which accounts for about 33% of the distribution.

Students answered the first several questions in a manner consistent with the concept of sample representativeness, focussing on the samples of 4 green and 2 reds that mirrored the population proportion of 2G:1R. All twelve of the students answered "4" to the question of how many green Gummy Bears might be in their packet. No student answered in a way that indicated a probabilistic solution. Instead, their explanations indicated that they regarded the question as a ratio problem.

When asked if every kid's packet would contain four green Gummy Bears, all of the students responded that there would be variation among samples, but some answers betrayed a resistance to thinking probabilistically. For some, this took the form of evoking a mechanism to explain the existence of non-representative samples. Others

saw the existence of non-representative samples as "imperfections" in the process of making up packets, which, if it had been done "right", would have given every student four greens and two reds.

Even students who were comfortable with the concept of variability, however, had incorrect ideas about the relative number of packets containing different combinations of green and red Gummy Bears. Their initial estimates of how many kids out of 100 would receive packets containing 4 green and 2 red Gummy Bears provide more evidence for the power of representativeness in this problem. Most students were convinced that a majority of samples would be representative of the population. While only one student believed that every sample would look the same, every student but one estimated that at least half of the packets would look the same.

In the next part of the interview we asked students to generate all possible combinations of green and red Gummy Bears. We originally thought of this request simply as a prelude to having them generate the entire distribution of packets, but we discovered, to our surprise, that 8 of the 12 students had trouble listing the seven possible combinations for a packet of candies. With the possible combinations listed (albeit sometimes painfully), we asked students to estimate how many kids would receive packets in each category. Now students were faced with a potential conflict between their initial guess, heavily influenced by the representativeness heuristic, and the necessity to distribute 100 packets among seven categories. For those whose initial estimate had been 75 or 80, there was a dilemma: how could they spread such a small number of remaining packets (20 to 25) among six additional categories?

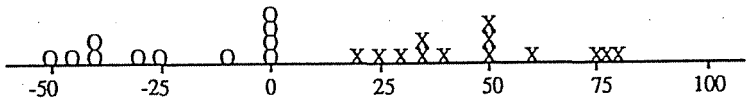

FIGURE 1

Student responses

(X : Final estimates: # in 100 with 4 green; O : Change from first to final estimate)

The general trend of students' responses to this dilemma can be seen in Figure 1, which shows the distribution of students' final estimates for the number of kids receiving packets with 4 green and 2 red Gummy Bears (X) and the distribution of changes from their initial estimates (O). The most obvious pattern that emerges from these data is that 7 of the 12 students lowered their estimates - in some cases by as much as 40 or 50% - when faced with the distribution problem. However, even after specifying the entire distribution and lowering their estimates of the frequency of the modal category, half of the students still significantly overestimated the frequency of the representative packet.

When distributing the packets among categories, students predictably retained their correct intuition that the 4G,2R packets would be most frequent. No student's distribution contained a peak at a point other than 4G,2R, and only three students constructed distributions in which a second category was tied with the peak. So their intuitions about the shape of the distribution were correct - but their estimate of the size

of the peak was, in general, significantly higher than the underlying model.

In general, in the Gummy Bears problem, students appeared to be unduly swayed by the implications of sample *representativeness* in constructing a distribution. They underestimated the frequency of samples near the tails of the distribution and overestimated the frequency of the modal sample, even when they were aware of the number of categories among which they had to spread their packets.

3. Runners : it's hard to get it right

The second problem asked students to evaluate two different ways of dividing up 400 runners - 200 fast and 200 slow - into blue and red teams. One way was to determine teams by the running ability of each runner, so as to make the teams as even as possible. The other was to assign runners to each team randomly, by choosing names out of a hat and assigning alternate runners to each team. The questions probed students' assessments of the "fairness" of the hat method, i.e. how likely it was to produce teams that were balanced in terms of fast and slow runners.

This problem provides a look at the influence of the concept of sample *variability* on students' thinking, as opposed to sample representativeness. After describing the two methods, we asked students to show us "how many fast and slow runners you think you would get on each team using the two methods. You can assume there are 200 fast and 200 slow runners." Without exception, students reported that the straightforward assignment of runners to teams based on their running ability would result in evenly-matched teams, with 100 fast and 100 slow runners on each of the two teams. With the hat method, however, many students reported that grossly unequal teams were possible. While several students claimed that the hat method could produce a fair assignment, most thought that teams with 150 fast runners and 50 slow ones were also possible outcomes of using the hat.

Fast	100	102	110	115	120	125	130	140	150	150	150	150
Slow	100	98	90	85	80	75	70	60	50	50	50	50

The above table illustrates students' estimates of how many fast and slow runners would end up on the red team. Eight of students' twelve estimates of team composition are very unlikely to occur. All of these eight are likely to occur less frequently than 1 in 100 samples; the seven most uneven samples are likely to occur less than 1 in 1000 times. Yet, some students were adament about how unfair the hat method would be.

Students' answers to the runners' problem contrast with their answers to the Gummy Bears problem in an interesting way. In that problem, they emphasised the likelihood of the representative sample - 4 green and 2 red Gummy Bears. A similar strategy in this problem would influence them to answer that the fast and slow runners would be evenly divided among the two teams. But no student proposed that answer. Since the representative sample *is* the most likely (although nowhere near as frequent as the representative sample in the Gummy Bears problem), it is interesting to speculate which of the contrasting characteristics of the two problems were most salient in

students' conceptions. One possibility is that the number of possibilities in the runners problem is so great that students feel the "accurate" answer (100 slow, 100 fast) is unlikely. Indeed, the frequency of the modal category in this distribution is only about 14%, compared to 33% in the Gummy Bears distribution.

Another possible explanation for the difference is students' experience with random assignment in situations of small samples. One student makes this quite clear:

> S2: "Well, usually, we do that in gym so you know, you go there, you go there, and they come out pretty uneven. One team is much better than the other team. It doesn't really work out very well."

Since this student (and others) does not recognise the effect of sample size, it may be that she generalises inappropriately from her gym class experience with small samples to the runners problem with samples of size 200.

The finite population in the running problem is another possible explanation for the different strategies that students adopted. It seems much less likely that a population of 400 will "split exactly" than that a small handful of Gummy Bears drawn from an enormous vat will reflect the population.

4. Interpretation and perspectives

Students' answers indicate that they lack experience thinking in terms of a *distribution* of samples generated from a particular population. Instead, they use heuristics - including, but not limited to, representativeness - to judge the likelihood of a particular sample. This heuristic-based thinking leads them to analyse different situations with similar underlying mathematical models from quite different perspectives. Thus, their answers in different problem settings fall in varying amounts under the influence of intuitions about sample representativeness or sample variability. This is not just a fact about aggregate student behaviour; in most cases, the same student answered the Gummy Bears question in a way consistent with sample representativeness and the Runners questions in a way consistent with sample variability.

There is some evidence in our interviews that the concept of "correctness", so prevalent in mathematics classes (and in school in general), may converge with students' tendencies to believe in sample representativeness. In the Gummy Bears problem, in particular, students' comments implied a tendency to regard the representative sample as the "correct" one. For example, one student commented that "the ratio is two to one so if you figure it out *exactly* that's what it would be *exactly*. And so I figured most people would get that." (Emphasis on tape.) Another student discussed the impossibility of "getting it right every time" in explaining why every sample would not necessarily look like the population.

Additional evidence comes from some open-ended interviews we conducted with students who had completed a one-semester statistics course. These suggest that the emphasis on a correct, accurate answer in mathematics class may combine with students' natural tendency toward the representativeness heuristic to produce a conception in which the representative sample is the one you get if you sample correctly. In this view, randomness is not sufficient to explain sampling variability - some mechanism or bias

must be postulated to explain it.

What we may be seeing here is an unfortunate collusion of the misconceptions students bring to statistics instruction and common patterns in school mathematics. Students start with the notion that samples are likely to be representative of the population a vast majority of the time; at times they even refer to such samples as "accurate". At the same time, the emphasis in school mathematics is (mistakenly) on accuracy, correctness, and lack of error. These two tendencies reinforce one another to undermine students' grasp of sampling variability. While they may be able to reproduce a sampling distribution in a structured problem, when faced with a more open-ended situation involving statistical inference, students often slip back into their notions of "accurate" samples, free of error. Even the use of the word "error" in the context of statistics is likely to be problematic, since it means random variation, a concept unrelated to the everyday use of the word in school.

5. Conclusion

The challenge to students learning statistical inference is significant. Their experiences in the world lead them to rely unduly on notions of sample representativeness and sample variability without unifying them into a single model of a distribution. Patterns of teaching in school mathematics may interact with the use of these heuristics to render students less flexible in their understanding of distributions. Even the vocabulary of statistics may make the subject matter more difficult to master.

Besides understanding better the structure of students' thinking about sampling and statistical inference, we need to evaluate different approaches to teaching the subject matter. We know that the way most of us were taught statistics was not pleasant - and was, in many cases, only minimally effective. We need to discover techniques that will help students keep what is useful of their prior intuitions and modify what is fuzzy to provide a solid basis for being a sophisticated consumer of statistical information.

Acknowledgement

This work was supported by the National Science Foundation, Grant MDR-8751893 and was carried out at Bolt, Beranek and Newman, Cambridge, Massachusetts. Opinions expressed here are those of the authors and not necessarily those of the Foundation. Reprints and full text of interview questions can be obtained from Andee Rubin, TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140, USA.