STUDENTS' STATISTICAL REASONING ABOUT DISTRIBUTION ACROSS GRADE LEVELS: A LOOK FROM MIDDLE SCHOOL THROUGH GRADUATE SCHOOL

Jennifer Noll[1], Michael Shaughnessy[1] and Matthew Ciancetta[2]
[1]Portland State University, United States of America
[2]California State University – Chico, United States of America
noll@pdx.edu

This paper provides a synthesis of the findings from three studies that investigated students' statistical reasoning about distributions of data and sampling distributions. Each study presented some of the same open-ended statistical tasks to students from different populations: middle school, high school, undergraduate or graduate students. The authors observed that students across all grade levels experienced difficulty in coordinating multiple aspects of a distribution. We will discuss two of the significant obstacles observed: lack of proportional reasoning skills when comparing different distributions, and difficulties in managing the natural tension between sampling variability and sample representativeness. Our research findings suggest that students throughout the grade levels need more opportunities to reason about empirical sampling distributions.

INTRODUCTION

Quantitative information has become omnipresent in our society. Numerical information is used to make data-based claims in sports, medicine, advertising, and political debates. As a result of these changes, statistics as a discipline is a necessary and important component of general education (see, Garfield & Ben-Zvi, 2008). The importance of statistics for all citizens has prompted concerns among educators and statisticians about the statistical education of both K-12 and college students, as well as the statistical preparation of teachers. These concerns have precipitated action by the National Council of Teachers of Mathematics (NCTM) and the American Statistical Association (ASA). In 2007, the ASA endorsed Guidelines for Assessment and Instruction in Statistics Education (GAISE). The GAISE report represents a substantial effort by statistics educators to provide guidelines and standards for statistics education from elementary school through college. These guidelines stress that students of statistics need to understand: (1) the important role context plays in statistics; (2) the role of variability in statistics; (3) different sources of variability; (4) how to reason informally about distributions of data and make sound data-based informal inferences; and, in general, (5) how to formulate questions, collect data, analyze data and interpret results based on an understanding of (1)–(4).

The three studies discussed in this paper investigated K-12, undergraduate and graduate students reasoning about distributions of data and empirical sampling distributions. The tasks used in these studies are consistent with the types of tasks suggested by Zieffler and colleagues (Zieffler et al., 2008), which support the development of informal inferential reasoning and relate to guidelines (2) and (4) in the preceding paragraph. The tasks were open-ended, non-routine and designed to assess student reasoning about empirical distributions of data, including student attention to measures of center, shape, and variability within and across distributions of data. Many students tended to base their conclusions about the data through a consideration of a single aspect of distribution, rather than a coordination of multiple aspects of the distribution. Across the three studies, we observed two significant cognitive obstacles experienced by the students: (1) applying proportional reasoning skills when comparing distributions of data, and (2) managing the natural tension between sampling variability and sample representativeness when making decisions based on empirical sampling distributions (similar results were found in Watson & Kelly, 2004).

This paper briefly describes the three studies, and then we present an analysis and discussion of one task given to participants from all three studies, including some trends in student reasoning we observed across all the grade levels.

BACKGROUND AND METHODOLOGY

The Three Studies

The first study was part of an NSF funded grant at a large urban university in the Pacific Northwest. The project participants included ten mathematics classes from two middle and four

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

high schools, with matched research and comparison classes. Data sources for the project included two large-scale class wide surveys, three semi-structured task based interviews, and three designed weeklong classroom teaching episodes. Comparison classes were obtained from the same school sites as the research classes, but did not participate in the teaching episodes. The task reported here was used in the second survey (n=236). Prior to administering the survey, the students participated in a weeklong teaching episode, conducted by the researchers, where the students engaged in activities with known and unknown populations that included sampling, predicting and reasoning about samples and sampling distributions (see, Shaughnessy et al., 2004).

The second study investigated graduate teaching assistants' (TAs) statistical content knowledge (see Noll, 2007). This study focused on how TAs' reasoned about empirical sampling distributions. The data corpus for this study consisted of a web based survey given to TAs at 18 universities across the United States (n= 68) and a series of semi-structured interviews with five participants selected from the survey population. Selection of the interview participants was based on availability, location, and the requirement that they had previous experience teaching an introductory college statistics course. The participants in this study comprise a volunteer sample.

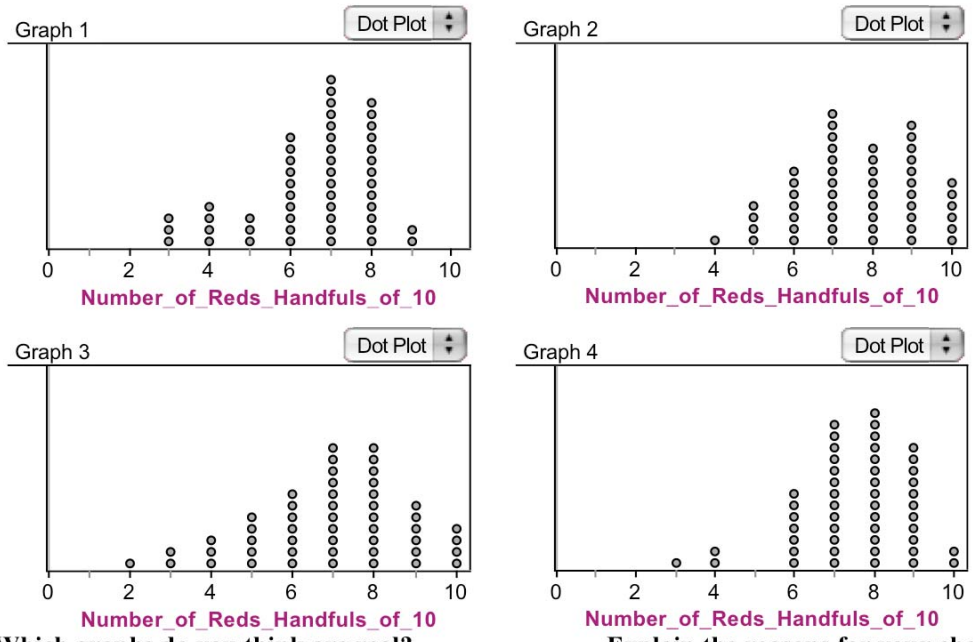
The third study investigated undergraduate students thinking about distributions of data and how they reasoned about empirical sampling distributions (see Ciancetta & Noll, 2006). Task based surveys were administered to undergraduate students (n=45) who were nearing the completion of their first or second term of an introductory college statistics course at a large research university in the Pacific Northwest. Thus, the students in each study (with the exception of the comparison students in the first study) had experiences in reasoning about sampling distributions. The participants in the third study also comprise a volunteer sample. The data collected in these three studies was primarily qualitative, although some quantitative methods were also employed.

The Real/Fake Task

The Real/Fake Task (shown in Figure 1) was presented to all students in all three studies. It requires students to compare and contrast four empirical sampling distributions and infer which distributions are more likely to come from the a priori known population described.

Figure 1.

A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 reds and 250 yellows, and graphed the number of reds. However, in this class some of the groups 'cheated' and did not really do the experiment, they just made up a graph. Here are some of the students' graphs from that class.



A) Which graphs do you think are real? _____Explain the reasons for your choices.

B) Which graphs do think are made-up? _____Explain the reasons for your choices.

The primary purpose of the Real/Fake Task was to investigate ways in which students utilize statistical thinking when reasoning about experimental sampling distributions. In particular, the researchers in each of these studies were interested in answering the following questions:

- What aspects of the distribution would students attend to (shape, center, variability)?
- What types of sampling distributions, if any, would students conceive of as unusual for this situation? How narrow or wide could a distribution be before a student considered it unusual?
- How much variability would students expect within and between sampling distributions?
- Would students expect the shape of the graph to be smooth or would they expect jumps in the frequency from one outcome to the next?

Success on determining which graphs are real and which are made up requires students to coordinate center, spread and shape of a distribution, as well as manage expectations for variability within and between sampling distributions. The researchers made up Graphs 1 and 3 ('Fake') and used a computer simulation to generate Graphs 2 and 4 ('Real'). Graph 1 was designed with an appropriate range, but shifted to the left. Both of Graphs 1 and 3 have an unusually high number of outcomes at four and below; for example, the probability of obtaining six or more handfuls (samples) of ten candies, of which 0–4 are red, is 0.0004. Graph 1 also has too few outcomes at nine and ten candies; for example, the probability of obtaining 2 or fewer handfuls (samples) of ten candies, of which nine or ten are red, is 0.00013. Lastly, Graph 3 not only had a range that was much too wide but was also designed to have a 'smooth' distribution in terms of the frequencies of outcomes, or heights, when reading the graph from left to right.

ANALYSIS OF STUDENT REASONING ON THE REAL/FAKE TASK

Summary Findings

Table 1 provides a summary of the number of correct identifications made by students from across the four grade levels from all three studies. The percentages of 2, 3, or 4 correct identifications in the real-fake task are fairly consistent across the grade levels (with the exception of the percentage of undergraduates making 2 and 4 correct identifications). More importantly, we wanted to understand why students made their choices. The remainder of this section discusses students' reasons for their choices and the similarity in types of reasoning across the grade levels.

Table 1. Total Number of Correct Identifications Broken Down by Grade Level

Number of Correct Identifications	MS comparison Students (n=25)	MS research Students (n=51)	HS Comparison Students (n= 71)	HS Research Students (n= 89)	Undergraduate Students (n= 45)	Graduate Students (n=68)
0	4 (16%)	3 (6%)	8 (11%)	6 (7%)	5 (11%)	3 (4%)
1	3 (12%)	1 (2%)	5 (7%)	11 (12%)	9 (20%)	4 (6%)
2	9 (36%)	20 (39%)	29 (41%)	33 (37%)	22 (49%)	24 (35%)
3	4 (16%)	9 (18%)	10 (14%)	12 (13%)	6 (13%)	18 (26%)
4	5 (20%)	18 (35%)	19 (27%)	27 (30%)	3 (7%)	19 (28%)

Table 2 summarizes the reasoning that students used to explain their decisions for which distributions were real and which were made up. Students' reasoning predominantly fell into four types: *attention to shape*, *attention to the tails*, *attention to spreads*, and *attention to centers*, roughly in that order of frequency. Also notable was that some students reasoned in Non-Appropriate (NA) ways that included not providing reasons, off task or very contradictory responses. When students coordinated two or more aspects of a graph they were coded as distributional reasoners (≥ 2 Aspects). Some students focused their reasoning on different aspects from graph to graph (e.g., center on one graph and spread on another graph), while others coordinated two or more aspects of the distribution (e.g., center and spread) on the same graph. In fact, most students that reasoned using two or more aspects of the graph focused on shape and tails. A few of the graduate students applied a formal procedure for determining the likelihood of a particular event. The analysis and discussion below focuses on students' attention to shape and to the tails, as these two forms of reasoning appeared significant and prevalent across grade levels. In

addition, our discussion focuses on students' common use of language across grade levels for describing shape and variability in the distribution.

Table 2. Students' Reasoning Across Grade Levels

Predominant Reasoning	MS comparison (n=25)	MS research (n=51)	HS Comparison (n= 71)	HS Research (n= 89)	Undergrad. (n= 45)	Graduate (n=68)
NA	14 (56%)	5 (9.8%)	19 (26.8%)	26 (29.2%)	8 (17.8%)	4 (5.8%)
Centers	0 (0%)	4 (7.8%)	3 (4.2%)	1 (1.1%)	2 (4.4%)	2 (2.9%)
Shapes	4 (16%)	19 (37.3%)	30 (42.3%)	25 (28.1%)	23 (51.1%)	33 (48.5%)
Spreads	1 (4%)	7 (13.7%)	2 (2.8%)	11 (12.4%)	0 (0%)	0 (0%)
Tails	6 (24%)	13 (25.5%)	7 (9.9%)	14 (15.7%)	2 (4.4%)	2 (2.9%)
$\geq$ 2 Aspects	0 (0%)	3 (5.9%)	10 (14.1)	12 (13.5%)	10 (22.2%)	27 (39.7%)

Attention to Shape

Graph 3 appeared to be the easiest for most students to discern that it was one of the made up graphs. Of the middle and high school students surveyed approximately 33% attended to Graph 3 first, and said things like Graph 3 "looked the fakest", "was definitely made up", "was cheesy fake", "was the least real looking", or "was too perfect". In the interview excerpt below, Molly (a 7th grade student) clarifies why she believed Graph 3 "was cheesy fake". Molly: *3 looks made up because, I know when I'm making graphs you sort of start with a really small one and then you get bigger–not necessarily the same amount. But, like this one, it goes up by one, up by one, then up by two, then up by two, then up four, then it stays, and then it goes down five, and then down by two.*

Many undergraduate students made similar comments for Graph 3, saying the graph "has too nice of a curve to it" and it has "too smooth of a distribution". Likewise, many of the graduate students reasoned by shape on the Real/Fake Task, especially in their descriptions of why Graph 3 was likely to be fake. The two excerpts that follow provide examples of graduate student justifications for why Graph 3 was made up. (Excerpt 1) Washington: *Graphs 3 and 4 are likely made up because the observations are "too perfect"; in such a small sample we would expect to see larger deviations from the true distribution.* (Excerpt 1) Interviewer: *When you say Graph 3 is too perfect, what do you mean by too perfect?* Amanda: *I expect in a real graph that they're not going to be in order 0, 1, 2, 3, 4 etcetera. Some are going to be higher than the one next to them and some are going to be lower.* Interviewer: *So the even steps up in frequency?* Amanda: *Yes, it goes up very smoothly. It doesn't have any dips in terms of 0* [red candies] *up to the 7, 8* [red candies] *and then from the 7, 8* [red candies] *it's decreasing back down. And that just strikes me as too perfect. Compare that to Graph 1 where we increase* [in frequency] *from 3 to 4* [red candies] *but then we decrease from 4 to 5* [red candies]. *So it* (Graph 3) *increases too smoothly and too evenly*.

The shape language used by students throughout the grade levels was similar and most of these shape arguments were used to argue that Graph 3 was made up. The data across grade levels provides some evidence that these students expected variability in real graphs in terms of variability in shape, but not necessarily in terms of statistical variability (i.e., spread or standard deviation). The responses of these students appear to be focused on their expectations for variability in the frequencies of the graph from one outcome to the next. For example, in the interviews both Amanda (graduate student) and Molly (a 7th grade student) elaborate that Graph 3 is "too perfect" because the graph does not have dips and bumps in its heights, which is something they expect of experimental data. Garfield and colleagues (Garfield et al., 2007) also found this belief among the students in their research.

Attention to the Tails

A second type of common argument students used regarding which graphs were real involved their attention to the tails of the distribution. Students attended to extreme values when determining whether or not Graph 1, 2, and/or 4 were real. Four categories emerged from student discussions of the extremes. Those categories are: Student expects more handfuls containing 10 red candies (More Highs, *MH*); Student expects fewer handfuls containing 10 red candies (Less Highs, *LH*); Student expects more handfuls containing between 0–4 red candies, (More Lows, *ML*);

Student expects fewer handfuls containing between 0–4 red candies (Less Lows, *LL*).

Reasoning about the extremes of the distribution is a potentially viable method for determining which graphs are real as long as that reasoning coordinates the ends of the distribution with the anchor of the population parameter. Since the parameter is 75% red candies, a person is more likely to pull out a handful of 10 red candies (i.e., *MH*) then handfuls containing 1, 2, 3 and/or 4 red candies (i.e., *LL*). When students reasoned in this way they tended to correctly identify the real versus fake graphs. However, when students did not coordinate the likelihood of the extremes with the population proportion they tended to think that getting a handful with all 10 red candies was much more extreme than a handful with 3 red candies, because they were not applying proportional reasoning from the population to the samples.

Many of the middle and high school students pointed out that there were a lot of handfuls of 10 reds in Graph 2 and/or that there were no handfuls of ten reds in Graph 1. Here, Jack (a 11th grade student) describes his reasoning regarding the number of handfuls containing 10 red candies in Graph 2 and Graph 1, and why Graph 2 should be made up and Graph 1 should be real. Jack (expects *LH*): *1 seems real to me because…there are no tens, because looking back at like when you were in our class, we never got one sample that was all one kind…and 2, it just seems like there are way too many tens for it be real.* Jack finds it reasonable that Graph 1 would have no handfuls of 10 reds because he views a handful of ten reds as somewhat extreme. In this excerpt Jack also recalls the sampling experiments done during the classroom teaching episodes, but he did not pay attention to the population parameters, and he did not notice that the population proportions in the Real/Fake Task were different than the one used during the classroom teaching episodes.

Many of the undergraduate students reasoned in a manner similar to Jack and the other middle and high school students from the first study. The following two survey excerpts illustrate undergraduate students' expectations for *LH* and *ML*. Both of these students argued that Graph 2 was fake since that distribution contained more handfuls with 10 red candies than they expected, and/or too few handfuls with 3 or fewer red candies. Jay (expects *LH*): *Based on the probability of red candy, graph # 1 and #4 seems like the most accurate distribution since the prob. of getting all 10 reds seems very small.* Sarah (expects *ML*): *It seems like some of the time 3 or less red could have been pulled out.* Although infrequent, there were a few undergraduate students who were able to coordinate their expectation for the ends of the distribution along with the actual population proportion. For example, Brie argued that Graph 1 was made up because she expected *MH*. Brie (expects *MH*): *There should be a few handfuls with all red, if 75% are red.*

Although the graduate students expressed a much deeper knowledge of the subtleties of sampling distributions, we nonetheless found that graduate students, like their younger counterparts, also experienced tension in balancing their knowledge of theoretical distributions with their expectation for empirical distributions. Graduate students can have similar gut intuitions for the extremes of the distribution as those expressed by the middle, high school and undergraduate students. The excerpt below provides an illustration of a graduate student who intuitively felt that getting a handful with 10 red candies was more extreme than a handful with 3 red candies. Despite the fact that this graduate student was able to reason distributionally in more formal contexts, she was unable to coordinate the tails of the distribution with the population parameter in this particular problem. Amanda (expects *LH* and *ML*): *I'm expecting to see something down here* [referring to the 2, 3 and 4 red candies spots on Graph 2]. *Especially taken in conjunction with the fact that I have, how many are piled here on 9? A lot. 11 in the 9 slots on Graph 2, and 6 in the 10 slot. And I feel like this is a little disproportionate. I've got nothing here* [in 2's, 3's and 4's] *and a lot going on at 9 and 10. And I would feel more comfortable. Watch this. This is just going to be awful. If I removed some off 9 and 10 and moved them over here* [to the 2, 3 and 4 red candy slots] *so that it looked more like Graph 3. The theoretical one that I think is implausible…. I'm having a battle in my head about theoretically what I expect to happen, which would look like Graph 3, and reasonably in practice what I have seen happen.*

CONCLUSION

Overall, students across grade levels had a tendency to rely mainly on the shape or the tails of the distribution when determining whether a graph was real or made up. Shape arguments were particularly prevalent for reasoning about Graph 3 and either shape or the tails of the distribution were prevalent arguments for reasoning on Graphs 1, 2 and 4. There are two compelling themes in

student reasoning that appeared across grade levels. First, students' had strong expectations that empirical sampling distributions have "ups and downs" in the frequencies, but seemed less attuned to statistical variation (e.g., spread, standard deviation). Students did not expect to see a graph that smoothly increased toward the center and then smoothly decreased after that. As a result, many students identified Graph 3 as a made up graph. Of course, this is an appropriate way to reason about graphs of empirical sampling distributions. However, students' focus on variability in the heights of the graph at different outcomes at the expense of any focus on the spread of the distribution made it difficult to identify fraudulent graphs that were not "too perfect". Second, it did not appear to be natural for students, even for many graduate students, to carefully consider the population proportion in coordination with their attention to the extremes of the distribution. As a result, many students tended to think that a handful of 10 red candies is much more extreme than a handful of 2 or 3 red candies, even though P(R) = 0.75. It may be that students just have a harder time envisioning a handful with all reds or all yellows than a handful that might contain a red or two and the rest yellows or vice versa.

Although these three studies have certain limitations (e.g., volunteer samples, or relatively small samples), the results presented here suggest interesting areas for further investigation and for pedagogical innovations. In particular, an obvious next step for further research would be to conduct a long term design research study that investigates how students construct knowledge of sampling distributions as they engage in principled instruction where they are asked to compare empirical sampling distributions from different population proportions along with explicit classroom discourse around the coordination of population parameters and extremes of distributions.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., & Witmer, J. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) college report*. Alexandria, VA: American Statistical Association. (Also available at http://www.amstat.org/).

Ciancetta, M. & Noll, J. (2006). Undergraduate students' difficulties assessing empirical sampling distributions. In A. Rossman & B. Chance (Eds.), *Proceedings of the 7th International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Garfield, J. B., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett & P. Shah (Eds.), Thinking with Data (Proceedings of the 33rd Carnegie Symposium on Cognition). (pp.117-147). New York: Erlbaum.

Noll, J. A. (2007). Graduate teaching assistants' statistical knowledge for teaching. Unpublished Doctoral Dissertation, Portland State University. Portland, OR. Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). Types of student reasoning on sampling tasks. In M. Johnsen Hoines & A. Berit Fuglestad (Eds.). *Proceedings of the 28th meeting of the International Group for Psychology and Mathematics Education*, Vol. 4 (pp. 177-184). Bergen, Norway: Bergen University College Press.

Watson, J. M., & Kelly, B. A. (2004). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics, and Technology Education, 4*, 371-396.

Zieffler, A., Garfield, J.B., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58.