INITIAL UNDERGRADUATE STUDENT UNDERSTANDING OF STATISTICAL SYMBOLS

Samuel A. Cook
Wheelock College — Scook@wheelock.edu

Tim Fukawa-Connelly
The University of New Hampshire — tim.fc@unh.edu

In this study we use the tradition of semiotics to motivate an exploration of the knowledge of, and facility with, the symbol system of statistics that students bring to university. We collected a sample of incoming mathematics majors in their first semester of study, prior to taking any statistics coursework, and engaged each in a task-based interview using a think-aloud protocol with questions designed to assess their fluency with basic concepts and symbols of statistics. Our findings include that students find symbols arbitrary and difficult to associate with the concepts. Second, that generally, no matter the amount of statistics that students took in high school, including Advanced Placement courses, they have relatively little recall of topics. Most can calculate the mean, median and mode, but they generally remember little beyond that. Finally, students have difficulty connecting practices or procedures to meaning.

Keywords: semiotics, symbols, statistics, baseline data, recall of K-12 schooling

Research Questions

There have been investigations of students' understanding of measures of center (Mayen, Diaz, Batanero, 2009; Watier, Lamontagne, & Chartier, 2011), variation (Peters, 2011; Watson, 2009; Zieffler & Garfield, 2009), and students' preconceptions of terms related to statistics (Kaplan, Fisher, & Rogness, 2009); however, relatively little work has focused on students' use of symbols (Kim, Fukawa-Connelly, Cook, 2012; Mayen, Diaz, Batanero, 2009). In this study, we attempt to explore students' understanding of the symbolic representation system in statistics from their secondary curriculum to establish the baseline of student knowledge of statistical symbols upon arrival at university. That is, we investigate:

1. What fluency with statistical symbols do undergraduate mathematics majors have upon arrival at university?
2. How does that vary based on prior course-taking?
3. How do students reason about statistical concepts, such as the normal distribution and sampling, upon their arrival at university?
4. How does that vary based on prior course-taking?

Initial results suggest that students, even those that had significant experience with statistics pre-college, struggle to recall the relationship between symbols and definitions.

Literature Review

Understanding symbolic representations of ideas is especially critical for mathematics students. Hewitt (1999, 2001a, 2001b) distinguished between "arbitrary" and "necessary" elements of the mathematical system. The aspects of a concept used by a community of practice are labeled "arbitrary," meaning they can only be learned through instruction and memorization. Those which can be learned through exploration and practice are labeled "necessary." Hewitt found that for students to become proficient at communicating with established members of the

community, they need to memorize the arbitrary elements and associate them correctly with appropriate understandings of the necessary elements. Hewitt noted that names, symbols and other aspects of a representation system are culturally agreed-upon conventions. For those who already understand them, the conventions seem sensible, but "names and labels can feel arbitrary for students, in the sense that there does not appear to be any reason why something has to be called that particular name. Indeed, there is no reason why something has to be given a particular name" (1999, p. 3).

The study of symbolic representations and the linkages between symbolic representations and the concepts themselves is at the heart of *semiotics*. Eco (1976) used the term "semiotic function" to describe the linkage between a text and its components and between the components. The semiotic function relates the antecedent (that which is being signified) and the consequent sign (which symbolizes the antecedent) (Noth, 1995). In the statistical community and the representation system in use within the community, there is a complex web of semiotic functions and shared concepts that "take into account the essentially relational nature of mathematics and generalize the notion of representation," and furthermore, "the role of representation is not totally undertaken by language (oral, written, gestures, …)" (Font, Godino, & D'Amore, 2007, p. 4). Throughout this paper, we recognize the inherent arbitrary nature of much of the symbolic system of statistics and draw on the notion of semiotic function as a means of linking a particular representation with the relevant concept.

This study is in line with the tradition of onto-semiotic research in mathematics education (Font, Godino, & D'Amore, 2007). It is situated in the context of statistics education, and designed to explore undergraduate students' understanding of the symbolic system of statistics at the start of university math major. It will attempt to describe the collection of semiotic functions the students had, focusing on those involving symbolic representations for the concept of the sampling distribution. It will also give a preliminary explanation of why the students constructed this particular set of semiotic functions by describing their understandings of the general representational system.

Methods

Data for this study was drawn from 7 participants in a mid-sized public university. All of the participants were first-semester students with a declared major in the Department of Mathematics and Statistics. We engaged each student in a task-based interview, during which the student completed a 14-item survey while using a think-aloud protocol. We created the survey by drawing on items used in a previous study on student understanding of statistics (Kim, Fukawa-Connelly, & Cook, 2012), adapting items from the *Assessment Resource Tools for Improving Statistical Thinking*, and developing our own tasks. Our goals were: to determine whether students could correctly recognize the symbols, to evaluate their understanding of what the symbols represented, and to evaluate their understanding of the related concepts (e.g., ideas related to the normal distribution).

All interviews were video-recorded and transcribed. We analyzed the transcribed interview data in combination with the students' written work using grounded theory (Strauss & Corbin, 1994). First, we made narrative comments indicating where students appeared to make claims about a particular statistical symbol or concept, and identified themes. Then, we collected the themes to create a set of initial codes and descriptions and came to agreement on the coding. Each author then read and coded all of the transcripts. We are currently developing profiles of the students and their overall proficiency with symbols. In coding, we described what students

knew and believed about symbols and concepts, as well as areas where they recognized that they lacked appropriate knowledge or held misconceptions.

Results

Students find symbols difficult to distinguish between and associate with related concepts

While previous research examined students' proficiency at the end of an undergraduate class, this study examined what students recalled from their high school years. For example, after Cam calculated a mean, he indicated its label was μ, then changed it to $\bar{x}$, then decided he did not know how to determine what it was. This finding is similar to that which was reported in (Kim, Fukawa-Connelly, & Cook, 2012). When compared with students who were currently enrolled in an undergraduate statistics class (Kim, Fukawa-Connelly & Cook, 2012), while the students in this study had slightly less proficiency with symbols, the difference was not great. The most common difference was that the students in this study had no ability to distinguish, and only limited ability to recall, the difference between symbols for populations and samples.

Students do not remember much at all

We interviewed students with a range of experience of statistics in high school, from no specific courses on statistics to two years of coursework. While the students with more years of experience typically knew that they had learned more about statistics and recognized more symbols as relating to statistics than those without fewer courses, they generally had no greater understanding of the concepts.

For example, Kenny had taken an AP Statistics course (but not the exam). When asked about symbols, he recognized many of them, but when asked to distinguish between a single individual being within one standard deviation and a sample of four being the same distance from the mean, Kenny claimed that they were equivalent situations. He also described a sample of four individuals with a mean as being "four people, each with the same score." Both of these conceptions are problematic; while one could be understood as a lack of sense of scale, the other is a fundamental misunderstanding of the mean as a measure of center. On the other hand, students with less experience in statistics exhibited more advanced reasoning (as articulated by Cook & Fukawa-Connelly, 2012).

Students have trouble connecting practice to conceptual reasoning

Students, even those who had taken pre-college statistics courses, showed an ability to calculate mean and median but had trouble understanding that these are tools used to determine the center of data. For example, when we asked Jeremy if mean and median describe similar things, he said they do not, even though he said they both generally described the middle of data. When we pointed out that he used the word "middle" to describe both a mean and a median, Jeremy acknowledged that he had, but even then reiterated that mean and median were not similar. At the conclusion of the interview, Jeremy said that he had learned how to do things in class, but never learned the concepts.

Jeremy also showed that he was able to apply a rule that he learned, but unable to conceptually expand on the rule. When asked to determine if a particular data point should be considered unusual in the context of normally distributed data, Jeremy referenced the "empirical rule." He said that because the point was within a standard deviation, it was not unusual. However, when asked the same question about the average of 4 random data points from the distribution, he said that the answer would not change because it was the same data.

Students generally have an ability to both calculate and give a verbal description of measures of center

All of the students, regardless of the amount of statistics that they have taken, can correctly determine the mean and median of a set of data. If they recall what the mode is, they are similarly able to determine it. When asked, they have each been able to give a verbal description. Similarly, while students are not typically able to remember the formula for, or how to calculate, the standard deviation, they are able to state that it is a "measure of variation." Jeremy described the standard deviation as a measure of "how concentrated they [the data] are around a certain value." James attempted to recall the formula for standard deviation and recalled that it "had a summation of *x*'s and a summation of *x-squares*" but was not sure of the exact formula. Considering that students typically, at most, calculate the standard deviation of a data set once or twice by hand, it is unsurprising that they retain no knowledge of how to do so.

Discussion

Students must know the symbols used by statisticians and associate them with their accepted statistical meanings. Students must also be able to distinguish statistics from parameters and to view statistics as variables when they are embedded in a certain context. The results show that, even though students have a variety of exposures to statistics prior to arrival at the university, their recall and facility with the concepts of statistics is still minimal. Generally, students have appropriate procedural knowledge for calculation of measures of center, but appear to have limited ability to describe the differences in uses of those measures. Although we need further research on this subject, our results suggest that it is best to treat students who have taken AP Statistics or other advanced coursework as having knowledge and beliefs similar to those who have *not* taken specific course on statistics. The major difference between these two groups seems to be that those who have never taken a statistics course are less confident in their abilities and are less likely to have misconceptions about concepts. We believe that our results have implications for both K-12 and collegiate instruction, suggesting that instruction should focus on developing conceptual understanding that can stay with the students over the long-term, rather than calculating procedures.

Questions for Discussion

1) What was surprising to you? Why?
2) Which of the findings appear to be the most interesting, from a research perspective? Why?
3) What other themes should be looked at, in addition to initial proficiency, in the data we have collected?

References :

Cook, S., & Fukawa-Connelly, T. (2012) Toward a Theory of Symbol Sense in Undergraduate Statistics. Proceedings of the SIGMAA RUME: Conference on Research in Undergraduate Mathematics Education. Portland, OR. http://sigmaa.maa.org/rume/crume2012

Eco, U. (1976). *A theory of semiotics.* Bloomington: Indiana University Press.

Font, V., Godino, J., & D'Amore, B. (2007). An onto-semiotic approach to representation in mathematics education. *For the Learning of Mathematics, 27*(2), 2-7, 14.

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

Hewitt, D. (1999). Arbitrary and necessary: Part 1, A way of viewing the mathematics curriculum. *For the Learning of Mathematics, 19*(3), 2-9.

Hewitt, D. (2001a). Arbitrary and necessary: Part 2, Assisting memory. *For the Learning of Mathematics, 21*(1), 44-51.

Hewitt, D. (2001b). Arbitrary and necessary: Part 3, Educating awareness. *For the Learning of Mathematics, 21*(2), 47-59.

Kaplan, J., Fisher, D., & Rogness, N. (2009). Lexical ambiguity in statistics: How students use and define the words: association, average, confidence, random and spread. *Journal of Statistics Education, 18*(1). Retrieved from http://www.amstat.org/publications/jse/v18n2/kaplan.pdf

Kim, H., Fukawa-Connelly, T. & Cook, S. (2012) Student Understanding of Symbols in Introductory Statistics Courses. Proceedings of the SIGMAA RUME: Conference on Research in Undergraduate Mathematics Education. Portland, OR. http://sigmaa.maa.org/rume/crume2012

Mayen, S., Diaz, C., & Batanero, C. (2009). Students' semiotic conflicts in the concept of median. *Statistics Education Research Journal, 8*(2), 74-93.

Noth, W. (1995). *Handbook of semiotics (Advances in semiotics).* Bloomington, IN: Indiana University Press.

Peters, S. (2011) Robust understanding of statistical variation. *Statistics Education Research Journal, 10*(1), 52-88.

Watson, J. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal, 8(*1), 32-61.

Watier, N., Lamontagne, C., & Chartier, S. (2011). What does the mean mean? *Journal of Statistics Education, 19*(2). Retrieved from http://www.amstat.org/publications/jse/v19n2/watier.pdf

Wenger, E. (2007). Communities of practice: A brief introduction. In *Communities of practice*. Retrieved from http://www.ewenger.com/theory/

Zieffler, A., & Garfield, J. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal, 8*(1), 7-31.