HOW STUDENTS LEARN ABOUT DATA DISTRIBUTION WHEN ADDRESSING A PROBLEM AFFECTING THEIR COMMUNITY

Yury M. Rojas
Department of Statistics and Mathematics, District University and University of the Andes, Colombia
yurymrojas@gmail.com

This research examines the way in which 9th graders learn about and come to understand data distribution and the related statistical notions of center and spread measures. In the study discussed herein the students in question identify and then collaborate to address the problems affecting their community. Using two sample groups, the students examine and make inferences about how the existing socioeconomic structure of their community affects nutrition. From the data gathered I find that cooperative work on a problem that is culturally relevant and meaningful to the students who propose it aids comprehension and recognition of data distribution and its measures.

Revisiting the literature on the learning of statistical notions, I found that normally students do not participate in contexts that permit them to develop their own questions in response to real data. In this way the students cannot construct conceptual tools with which to reason about data from its distribution and their supposed learning is confined to calculations using various statistical methods (Petrosino, Lehrer & Schauble, 2003). Cobb and McClain (2004) indicates that the difficulties that students have with reasoning from data are due to the fact that they frequently learn statistics as a collection of tools and techniques. Similarly, MEN (1998) holds that when children form questions about the physical world and try to explain the situations and problems that arise from it, they need to gather information, determine its importance, how to collect, represent, and interpret it in order to obtain answers, generating new hypotheses and experiments in the process. These activities, furthermore, allow students to leverage other portions of the curriculum and put in practice their knowledge of numbers, measurement, estimation and problem solving strategies.

The purpose of this study is to explore the way in which 9th grade students learn about data distribution and the related statistical notions of center and spread measures when they themselves identify and collaborate to address nutritional, implicitly socioeconomic, problems affecting their community. To address this, I administered pre- and post-experiment tests composed of four assesment scenarios that were slight modifications of the National Assessment of Education Progress (NAEP). I analyzed the data qualitatively to identify increases in understanding related to statistical notions, and quantitatively using the Wilcoxon test to examine for any significant differences between the averages. I also recorded and qualitatively analyzed several individual class sessions to better understand the way in which students' comprehension developed as instruction proceeded.

The motivation for this study comes from work that the 9th grade students completed as volunteer teachers at the Santa Maria Foundation, an educational foundation created for children from poor socioeconomic backgrounds that is financed by the volunteer teachers' own school. This latter institution serves students from better socioeconomic backgrounds than those of the Santa Maria students, allowing the experiment to compare results across the two schools. At the outset of the experiment the 9th grade volunteer teachers hypothesized that students of lower socioeconomic backgrounds had poorer nutrition than students of upper-middle class backgrounds because the former lacked sufficient funds to buy food enabling a balanced diet. To better address this issue the 9th grade volunteers recognized they needed further information and began to design and apply instruments, including polls and interviews, to obtain data with which to test their hypothesis. Data collection encompassed the number of child participants in the study, the socioeconomic status of participants' families, their age, height and body mass index. It should be noted, however, that the interest of the 9th grade volunteers in the experiment went beyond confirming their hypothesis to real comprehension of the studied phenomenon. After data collection they began to employ statistical notions that allowed them to describe the data, understand its behavioral patterns and interpret it in terms of the the identified variables and the study's participants.

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]

At the beginning of the experiment one of the groups of volunteer teachers only calculated and theoretically defined some of the statistical measures (arithmetic mean, mode, standard deviation) without establishing connections between them and the research question. Only later did they begin to conceptually link the two and demonstrate understanding that the experiment required more than algorithmic application of a formula. For example, upon analyzing the weights of the children that participated in the study, this group of volunteer teachers, more than merely applying the standard deviation formula, began to recognize characteristics of the data's dispersion. They affirmed a higher standard deviation corresponded to greater diversity among the reported weights of study participants and began to use words—"agglomerated," "concentrated," "grouped," and "dispersed"—that reflect comprehension of the data distribution.

Student 10: *The weight, we see that the averages are similar, 35.43 and 35, while the standard deviation is 7.39 in the school and 5.9152 in the foundation.*
Student 20: *This means the school children**, t**heir data are further from the mean.*
Student 10: *Scattered, the deviation measures the distance between data or the distance between data and the mean? Tell me.*
Student 4: *The deviation is the distance with respect to the mean.*
Teacher: *Is it the distance from one data point to the mean?*
Student 20: *No, it is the average of the distances from all data points to the mean.*
Student 20: *The average.*
Student 4: *I didn't say the average*
Student 10: *It's important because it isn`t one data point but all data*
Teacher: *So what does standard deviation tell you when you're doing analysis?*
Student 4: *Whether the data is grouped*
(*the student interrupts: or is dispersed****)***
Student 20: *For instance the data from the school*
Student 4: *are more scattered*
Student 20: *Are more scattered, further from the mean.*
Student 4: *They're not as concentrated as the children from the foundation.*
Student 20: *There is more variety between the girls weights, while in the foundation they are closer to one place……..ok, it isn´t a place.*
Student 10: *They are more similar.*
Student 4: *They are more agglomerated.*
Student 20: *Good.*
Student 10: *They're more concentrated.*
Student 20: *They're more concentrated on one side, while in the school there are more different weight measures and different sizes. That's what it means.*

The improvement of this group in its ability to understand the center and spread concept is also evident in the presentation of their final experiment results:

Student 10: *The school has a 7.39 standard deviation, while the foundation has 5.91.*
Public: *And the mean is the same?*
Group: *It's the same!!!!*
Student 4: *In this way, we can say the foundation's data is more grouped in one part and has fewer spreads, while the weight of school participants can vary (she moves her hand up and down to illustrate the fluctuations in the data) though it depends on the case. Then, we found one participant who weighs 40 kilos while another weighs 20.*
Student 4: *Only one person can affect all others …..*
Teacher: *Good, and in terms of all participants' weight, what can we say? Is the data more grouped or more dispersed?*
Student 10: *There is more variability between weights in the school.*
Student 4: *And that there are more heavy ones and others that are slimmer.*
Student 10: *This explains the variability, some can be….*
Student 4: *obese!!! And others scrawny!!!!*

Further, to reason about the spacial distribution of the data the student volunteers consider what a high standard deviation means in the context of participants' weight. They come to understand that this analytical tool measures data variability, and that a high standard deviation encompasses data points as disparate as the study participants weighing 20 and 40 kilograms, respectively. This leads them, then, to question what such data variability says about the population under study, specifically, that greater variability means there can be greater numbers of both heavier and skinnier students.

The findings of this research show that students demonstrate the greatest understanding when qualitatively describing data organization and behavior, and when identifying features including holes in the distribution, extreme values and outliers. These descriptions reflect a perception of variation in the data, which motivates the use of statistical measures allowing the students to describe and better understand the data set. Most important, however, is the finding that the students learned to use these statistical tools and concepts in the context of identifying and addressing the most pressing problems facing their own communities, a finding which, if corroborated by further research, could serve to amplify authentic learning in statistics classes.

The test employed allowed me to obtain information necessary to track the students` comprehension of the characteristics of data distribution, which advanced as the experiment progressed. Notably, I found that previously developed problems that provide the students with data and that do not stem from the students` realm of personal experience generate lesser innovation, i.e. use of algorithms, in the students´ use of statistical tools. They struggle to manage other types of learning, as the situations require the application of certain measures. In this regard Gravemeijer & Bakker (2004) suggests that some designed, simulated statistical problems allow students to put to use only one of the measures, which does not allow performance of alternative actions such as problem formulation, data collection, among others, which are tools that students must develop to achieve a better understanding of the variables involved.

It is important to have considered the potential to address real problems that pertain to the students´social context and the issues surrounding it. In a significant way the students participated in the entire process of problema resolution from planning their approach to analyzing the data. The potential for learning is enormous when one uses a problem that allows the students to become heavily involved in the resolution and interact with both qualitative and quantitative variables, for which the students must not only determine appropriate measures but also take into account factors such as data values and variable types.

When students approach their own questions they need to draw upon and develop prior learning to analyze the resulting data, in the process demonstrating statistical application and conceptual development which manifest as real use of statistics as a tool of understanding. Ben Zvi (2001) believes that they operate along four dimensions (research cycles), including the selection and definition of a real problem, construction of a plan to address it, data collection, and analysis to draw conclusions. These are types of thought that are related to problem-solving strategies and interrogative cycles and provisions such as imagination and openness to change perceptions. I find that when my students worked on a real problem in collaborative teams, they developed an authentic manner of problem solving in accord with the dimensions proposed by Ben-Zvi (2001), which relates to the real performance of a statistician.

Finally, although the students generated innovative insights about the characteristics of center and dispersion, they made little progress regarding the shape of the distribution and did not consider potential biases in the data. The scatter plots constructed by the students allowed them to see aspects of the data´s organization, density and dispersion, but I consider work with different types of representations, such as histograms and bar graphs that generate discussions about the shape of the distribution, to be relevant for future study. Also, the results of my study produced some evidence that the characteristics of data sets (which have equal average, notable differences in the variation, etc.) may or may not facilitate recognition of certain characteristics of the distribution.

REFERENCES

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about Distribution. In *The challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147 - 168).

Ben-Zvi, D. (2001). Junior high school students` construction of global views of data and data representations. *Educational Studies in Mathematics.* The Netherlands: Kluwer. (pp. 35-65).

Cobb, P., & McClain, K. (2004). Principles for instructional design in developing statistical reasoning. En: *The challenge of developing statistical literacy, reasoning and thinking.* The Netherlands: Kluwer.

MEN (1998). *Lineamientos Curriculares.*

Petrosino, J. Lehrer, R., & Schauble, L. (2003). Structuring error and Experimental Variation as Distribution in the Fourth Grade. M*athematical thinking and Learning* (pp. 131-156).