FOUR THEMES ABOUT CONSTRUCTION OF HISTOGRAMS

Megan Mocko and Dan Maxwell
University of Florida, USA
megan.mocko@warrington.ufl.edu

This article reports findings from a multi-year mixed-method study that explored the challenges students encountered while answering fundamental histogram questions. In the first stage in 2017, an exploratory study was conducted using think-aloud protocols. In the second stage in 2019, an online survey was developed to confirm the preliminary themes. Four themes were identified. The themes are as follows: (a) identify the correct type of graph to explore the center and variability of one quantitative variable; (b) realize data has its own expressed shape; (c) lose concrete identifiers; and (d) create bins and count individual values. These themes will add to the current knowledge of students' difficulty with interpreting and describing histograms. This knowledge can help teachers prepare activities to promote understanding of histograms.

INTRODUCTION

What appears simple to a seasoned statistics instructor, professional data scientist, or working statistician often proves to be an overwhelming challenge to the student enrolled in their first statistics course. Clearly, students must master both conceptual and constructional understandings to comprehend what a histogram represents. Hence, the goal of this article is to advance a way of thinking about histogram learning. The authors first interviewed students to reveal preliminary themes, and then conducted a second analyses to confirm these themes. The result of these investigations is four final themes of construction of histograms.

Before looking at these new findings, what are current understandings about histograms? Past research shows us the importance of histograms and their role in the introductory statistics classroom. The third learning goal in GAISE 2016 states, "Students should be able to produce graphical displays and numerical summaries and interpret what graphs do and do not reveal" (GAISE College Report ASA Revision Committee, 2016, p. 8). These graphs include histograms and bar graphs that are often covered within the first part of the semester in introductory statistics classes. Although these appear to be basic visualizations, these graphs confuse students, especially concerning the case value plot (Kaplan et al. 2014, Meletiou-Mavrothesis & Lee, 2010). A histogram is a graph that displays one quantitative variable, whereas a bar graph displays one categorical variable. For a histogram, each bar is a bin that is a range of quantitative variables. By contrast, a case value plot displays a bar for each observation, and the bar represents that value for that observation on the vertical axis.

Whitaker and Jacobbe (2017) found that K–12 students also had difficulties with histograms. After the researchers shared their findings, they close with a call for additional qualitative research on this topic by writing:

> Despite the multi-year, iterative process to design these items, the resulting data are still limited. While the distractors for each item were carefully chosen to enable data to be collected on specific misunderstandings and the constructed-response items allowed students to respond using their own words, these items do not allow researchers to capture the type of rich data that would come from cognitive interviews and think-aloud interviews as they progress through school and mature in statistical thinking. (p. 101)

METHOD, DATA COLLECTION, & DATA ANALYSIS

To acquire this experiential understanding, the researchers directed the students to work through the study instrument by thinking aloud. This method of soliciting information of student understanding is described in Schoenfeld (1985).

There were two studies—an initial exploratory study conducted in 2017, followed by a confirmation study in 2019 (Link to questions: https://docs.google.com/document/d/1oc7p5yJQw-ZLB-jNQCuC1NM5Zh71GXYrwEj_nTkk5GU/edit?usp=sharing). The 2017 study instrument consisted of six questions. Questions one through three had the students distinguish and focus on the concept of what made a histogram. Questions four through six had them focus on the horizontal axis and vertical axis. Question three was described in a presentation by Kaplan et al. (2014). Question six came from the

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

Kaplan et al (2014) study, as did the inspiration for questions one and four. And finally, one of the authors wrote questions two and five in order to delve more deeply into what students consider to be a histogram.

For the 2017 stage, students in an introductory statistics course during the fall semester at a university in the United States' southeastern region were invited to participate in an interview. In total, fourteen interviews took place—all but two of these involved exchanges between two students. If one of the students failed to show up, the sole participant was asked to talk to one of the authors as if they were a fellow student.

After all of the 2017 interviews had been completed, the authors then graded each response as correct or incorrect and discussed participant explanations for each of their answers. Using this rubric, each investigator coded participant responses separately and then they met together to discuss each interview and arrive at a mutually agreed reasoning level and reasoning explanation for each question. The discussion continued until a consensus was reached.

After analyzing the data from the initial study, the authors then developed a survey to confirm the preliminary themes that emerged during the 2017 study. The 2019 survey consisted of 13 items and included three questions from the 2017 protocol. For the 2019 stage, students enrolled in an introductory statistics course during the summer and fall terms were asked to complete a survey. One of the survey questions asked if the student would be willing to complete a short interview, and a small subset of those who responded affirmatively was randomly selected to participate in this follow-up activity.

RESULTS

In this section, we will present each of the themes realized through the 2017 interview process. An examplar of the student's reasoning process and an analysis of the students' responses to the question and data from the confirmatory 2019 study will be given.

Identify the Correct Type of Graph to Explore Center and Variability for One Variable

After analyzing responses from the 2017 interviews, it was clear that many students lacked even the most basic histogram knowledge. For question 2, out of the 14 interviews conducted in 2017, only two responded correctly to this question, indicating that a histogram displays bars that represent how frequently a single quantitative variable appears in that range of values. In nine of the interviews, participants responded that a histogram visualizes a relationship between two variables. Interview 7 included the following reasoning about a histogram: "A histogram is a graph made to visually represent the relationship between two variables. The *x*-axis represents the explanatory variable while the *y*-axis represents the response variable." Interview 8 stated that, "a histogram is a graph used to show both quantitative and categorical data. An example would be a graph displaying utility bills for each month of the year." The following statement from interview 12 displays an accurate understanding of histograms: "A histogram is a graph that typically has just one independent variable that is associated with its certain frequency. Histograms look like bar graphs. However, their data is close together and is touching."

The 2019 survey that was given to more students in three different classes confirmed these initial findings. In particular, question 3 asked the students about what type of data would be used to look at amount of money spent on a vending maching. The percent correct was only 35%. There were four answer choices given, so the percent correct is only a little bit better than random chance.

Realize Data Has Its Own Expression

Another theme that was common was that data had its own expression. In other words, a histogram is not a correct graphic just because it shows data in a particular manner. That is, it is not up to the individual observing the data to impose a shape to it. For example, in question three of the 2017 study instrument, the students were given graphs of the number of Olympic medals won by countries that received at least one medal. For the problem, there are four options. Three of the four (A, B, and D) are case value plots that list the countries at the bottom with the height of each bar being the number of medals won by that country. For A, the countries are listed in alphabetic order, for B the countries are ordered such that the data assumes a bell shape, and for D the countries are ordered such that they assume a left skewed shape. Students in only four of the fourteen interviews selected option C (the histogram) as the correct answer. In interview 8, one student described why they felt that the answer should be the graph that was bell shaped as follows:

> And between D and B, this one is very left-skewed, I think, is right? The right term for it? But I think this one allows you to see the center better, because it groups the countries that won the most medals towards the middle, so you're going to be able to see who won the most medals of anybody. And it's bell-shaped, so it's pretty easy to describe the shape, and the spread would be the number of medals, I think. And so, maybe B would be the best graph to describe it.

Lose the Concrete Identifiers

In the process of building a histogram, the identifiers for each of the data points is lost or disconnected from its identifier. In the 2017 interviews, there were multiple cases in which students remained persistent in their desire that this information be retained. It is helpful to once again look at the responses to question 3 of the 2017 protocol. The students frequently cited a lack of identifiers as a reason for why a graph would not be a histogram. Interview 5 stated,

> I don't think that B would be one because it doesn't say anything about the countries. You don't know what you're comparing. It just has the rate and then the frequency. It doesn't have, 'Oh, which country are which.' So that's why I don't think B would be one.

These comments were categorized under the theme of students unwillingness to lose identifiers. Question 13 on the 2019 survey was designed to further probe student understanding of this step. Question 13 asks students, "When making a histogram, it is important to keep the name of the student associated with the price in the graph. True or False." For this question, 87% for the question correct. In Question 10, when the participants were asked to determine the horizontal and vertical axis the, highest marked answer item was "The x axis should be the months of the year and the y axis should be the number of text messages per that month." The correct answer would be for the student create the numeric bins and find the frequencies and thus lose the identifiers.

Create Bins and Count Individual Items

Correct histogram reasoning concludes with an accurate understanding of the binning process whereby frequency bars are constructed. After bins have been created, the individual observation values are now masked. Evidence for this reasoning stage could be found in the responses to question 4 of the 2017 protocol. The histogram had a space in which there were no observations, and thus, there wasn't a bar. When asked if a histogram was missing a year's worth of data, the participants in Interview 2 assumed that each bar represented one year, rather than the possibility that a bar contained the results of multiple years. Additionally, these students agreed that "there appears to be one year in which data wasn't recorded" by marking this statement as true on the paper protocol. They then wrote a note on the sheet that "multiple years are missing."

Question 4 on the 2019 survey was designed to assess student understanding of this fourth and final step of the model. For this question, students were asked directly to identify an essential step for creating a histogram. The only correct answer offered was creating bins. Students scored poorly on this question, with 54% answering question 4 correctly.

DISCUSSION

In this paper, an initial group of students was interviewed about their understanding of a histogram. During the analysis of these interviews, four themes emerged: (a) students not identifying the correct type of graph to be used for one variable, (b) students not realizing that data has its own shape, (c) students not learning that identifiers have to be lost, and (d) students not understanding the binning procedure. After these interviews were conducted, a confirmatory survey and follow-up interviews were conducted to confirm these findings. Thus, these four themes were established.

CONCLUSIONS

For students, the histogram presents a variety of reasoning challenges. Repeatedly, students clung to a desire to talk about a relationship between two variables, rather than talk about the variable given to them. They were reluctant to let go of the identifiers and thus unable to select the appropriate graph to answer the question of interest. They wanted the graph to be easier to understand and consequently wanted the histogram to be in a shape that would make it easy to understand. The binning process created a unique set of difficulties as well.

Implications for Instruction

Instructors need to be aware of the processes that underpin correct histogram thinking and provide pedagogical support where needed. Here are four suggestions for teaching. (a) Discuss the kinds of graphs one should use to explore certain types of investigative questions. For example, a histogram is a good candidate for questions that revolve around understanding center, shape, spread, and unusual points for only one quantitative variable; however, a scatterplot is a good candidate for questions that revolve around understanding the relationship between two quantitative variables. (b) Emphasize that data comes with its own shape, that the maker of the graph does not impose this shape. Show as many different shapes as possible, so that it cannot be assumed that the normal distribution is the result of all data investigations. (c) Demonstrate the process of removing observation labels as the data is brought into the histogram. (d) Ask questions that make students realize that the data label is gone and discuss the ramifications of this process. Do not assume that students easily see the transition from the raw data to the binned items in the histogram.

Directions for Future Research

A goal of any instructor should be to improve student understanding of histograms. So, a natural starting point is this: "What interventions are most effective in helping students arrive at a correct understanding of histograms?" For example, is explicit commentary that histograms describe only one quantitative variable sufficient for constructing correct histogram reasoning processes? Or do students need additional scaffolding? What level of intervention is needed to bring students to a coherent level of understanding?

REFERENCES

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education college report 2016.* American Statistical Association. https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports

Kaplan, J., Gabrosek, J., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education, 22*(2). https://doi.org/10.1080/10691898.2014.11889701

Meletiou-Mavrotheris, M., & Lee, C. (2010). Investigating college-level introductory statistics students' prior knowledge of graphing. *Canadian Journal of Science, Mathematics, and Technology Education, 10*(4), 1279–1303. https://doi.org/10.1080/14926156.2010.524964

Schoenfeld, A. (1985). Making sense of "out loud" problem-solving protocols. *The Journal of Mathematical Behavior, 4*(2), 171–191.

Whitaker, D., & Jacobbe, T. (2017). Students' understanding of bar graphs and histograms: Results from the LOCUS Assessments. *Journal of Statistics Education*, *25*(2), 90–102. https://doi.org/10.1080/10691898.2017.1321974