REASONING ABOUT ASSOCIATION FOR CATEGORICAL DATA USING CONTINGENCY TABLES AND MOSAIC PLOTS

Sheri E. Johnson, Ph.D.
University of Georgia
sheri.johnson25@uga.edu

With the implementation of Common Core in most states, the pre-k-12 mathematics curriculum now contains a significant amount of probability and statistics, mainly situated in middle and secondary grades. Statistical association is a challenging concept, and secondary students are expected to use contingency tables to begin to reason about association of categorical variables. This requires proportional reasoning, which is a focus of middle grades mathematics and necessary for more advanced study but remains a struggle for even most adults. Researchers call for use of multiple representations to develop conceptual understanding, and I consider a traditional contingency table in addition to a mosaic plot to see how students reason through a series of tasks.

Keywords: Cognition, Data Analysis and Statistics, Representations and Visualization, Geometry and Geometrical and Spatial Thinking

Statistical skills develop over time and in order for all high school graduates to have statistical literacy, instruction should begin early and expand through middle and high school (Bargagliotti et al., 2020). Association can exist between variables that are quantitative, like a person's height in centimeters, as well as variables that are categorical, like a person's eye color. Association of two categorical variables is included in the eighth and ninth grade curriculum in most states. Statistical (in)dependence can be determined numerically or visually and it certainly requires proportional reasoning, which researchers have identified as a "major connecting idea" when reasoning with probability and statistics where it is important to help students make explicit connections between data and proportions (Watson & Shaughnessy, 2004).

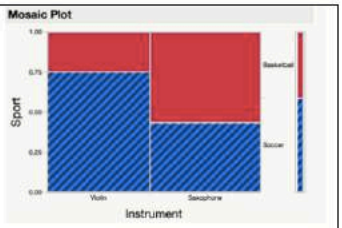
When considering bivariate data that are quantitative, there are well developed and standard graphical methods to aid students in determining (in)dependence, namely the use of a scatterplot (Friendly, 1999). The widely accepted Cartesian plane serves as a structure to visualize the data and determine if a linear or other type of association might be present. When the data are categorical, a two-way contingency table is used, but a standard graphical display does not exist for considering association of categorical variables (Friendly, 1999). Bar graphs, either segmented or side by side are often used to display frequencies in contingency tables; however, a mosaic plot is a default display used in some statistical software. Researchers in Australia (Pfannkuch & Budgett, 2017) recently noted some promising results of students working with an interactive mosaic plot, which assisted students in appropriately applying proportional reasoning with problems dealing with probability, especially when considering independence. A mosaic plot is based on a unit square that is vertically divided in proportion to marginal frequencies of one variable and further divided into rectangular regions that are proportional in area to each of the joint frequencies (see Figure 1).

Mosaic plots are often non-numerical and can be used to understand what the displayed data implies quantitatively and determine independence. Visualization can aid engagement with meanings and concepts that are not readily available through symbolic representation and when information is displayed visually, we are able to "see" the story, picture a cause-effect aspect of a relationship, and vividly remember it (Arcavi, 2003). Visuals can "group together clusters of information that can be apprehended at once" (Arcavi, 2003, p. 218), and "visualization at the service of problem solving, may also play a central role to inspire a whole solution, beyond the merely procedural." (p. 224).

In: Sacristán, A.I., Cortés-Zavala, J.C. & Ruiz-Arias, P.M. (Eds.). (2020). *Mathematics Education Across Cultures: Proceedings of the 42nd Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Mexico. Cinvestav / AMIUTEM / PME-NA. https:/doi.org/10.51272/pmena.42.2020

(a) Contingency table (Association)

Sport/Instrument	Violin	Saxophone	Total
Soccer	58	36	94
Basketball	13	47	60
Total	**71**	**83**	**154**



(b) Contingency table (Independence)

Sport/Instrument	Violin	Saxophone	Total
Soccer	43	51	94
Basketball	28	32	60
Total	**71**	**83**	**154**

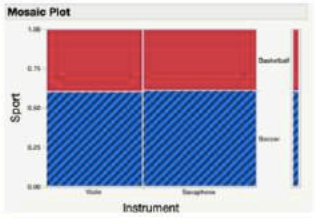


Figure 1. Contingency tables and corresponding mosaic plots.

Although elementary students are not likely to reason proportionally, previous studies indicate that younger students can reason correctly about association when only doubling and halving are required. The present study investigates how pre-k-12 students reason about association of categorical variables using contingency tables with and without mosaic plots.

Framework

Seminal work aimed to understand how students reason with complete contingency tables (Batanero, Estepa, Godino, & Green, 1996) provides the basis of my framework. Proportional reasoning requires the comparison of ratios and the use of all four cells in a multiplicative manner. I developed a framework which includes eight conceptions of reasoning with contingency tables (see Table 1), which are based on the five levels (L1-L5) identified by Perez-Echevarria (1990, as cited in Batanero et al., 1996).

Table 1 Graph Type and Variable Values for Different Items

Code	Description and features
N0	No cells in the table are used to decide about independence or association.
N1	No interior cells and one or more marginal cells are used to decide about independence or association.
L1	Localist-1: One interior cell in the table is used.
L2	Localist-2: Two cells in the table are used.
L3	Localist-3: Three cells in the table are used.
A1	Localist-4: All four cells in the table are used in an additive way.
P1	Proportional-1: All four cells in the table are used with proportional reasoning that compares risk (part to whole ratios). One conditional relative frequency is compared to another, focusing on the interior cells.
P2	Proportional-2: All four cells in the table are used with proportional reasoning that compares risk (part to whole ratios), and compares one conditional relative frequency to a marginal relative frequency.

P3	Proportional-3: All four cells in the table are used with proportional reasoning that compares odds (part to part ratios) and compares the odds for one category to another category for the same variable through subtraction or a ratio.

When considering the problems where the mosaic plots were provided, I used the same base codes and appended an additional code to indicate how the mosaic plot seemed to function. I considered whether the mosaic plot was a hindrance (M-), seemed to have no impact on the solution (M), or was helpful (M+).

Research Design

Since my interest is of the "how" and "why" nature, a qualitative, multiple-case study design is appropriate (Patton, 2005). I conducted think-aloud interviews (Charters, 2003) with seven participants that ranged in age from seven to 17 because I wanted to get a sense of ways that students across upper elementary, middle school and high school would respond to the same tasks. The reasoning about contingency tables of students in this age range has been underrepresented in past studies.

Each interview was semi-structured and used a protocol I develop based on the literature. The tasks all used the same context, but varied in the completeness of the contingency tables, numerical values of the frequencies, and whether there was an association among the variables. The words "association" and "independent" were used in the questions along with an additional explanation of their meaning. For the first part of the interview, I provided them with a series of problems with contingency tables and asked questions to ascertain their understanding of what the values in the tables represent. Then I introduced a mosaic plot by having them create one with a simple example. After verifying they could reason with it in conjunction with a contingency table, I presented two of the initial problems along with an accompanying mosaic plot. This study focuses on these two tasks. I concluded by asking them questions about the mosaic plot in general.

I recorded each interview with two video cameras, capturing both a close-up view of the student work as well as a broader view of the student to include gestures and facial expressions. Each of the seven interviews was transcribed and both an augmented transcript, noting participant actions, and a lesson graph, including notes of interesting moments, were created. The framework was used to code the data.

Results and Significance

The mosaic plot was never a distraction and was most often helpful. The two tasks considered in this study proved difficult for most students, and only two of the older students, Scott and Klaus, were able to provide comprehensive explanations for the tasks (Scott: Task 1 P2/P2M, Task 2 A1/P2M+; Klaus: Task 1 N)/P2M+, Task 2 A4/P2M+). These correct explanations occurred each time a mosaic plot accompanied the problem, but only once when it did not (Scott, Task 1). When the mosaic plot was provided, it improved all students' reasoning with the exception of this one problem Scott was able to solve without it.

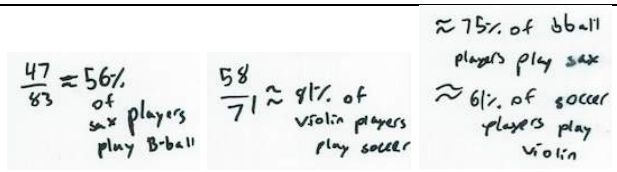
The younger students always showed improved reasoning when the mosaic plots were provided, but because of their limited proportional reasoning, they were not able to provide a completely correct solution and justify their reasoning. As the literature suggests, they were able to use numbers and benchmark fractions to reason about the data, but they conflated percentages with frequencies. Additionally, they did not have an understanding of the structure of a contingency table, often indicating the marginal frequencies for the rows and columns were a different group of people.

Some of the improvements in performance on the problem when the mosaic plot was included could be due to other factors like seeing the problem for a second time and working with different contingency tables between reasoning without and with the mosaic plot. However Scott and Klaus both verified they were looking at the mosaic plot and used it to solve the problem and overall, explanation and justification was more limited when there was no mosaic plot provided. Clear and complete explanations more frequently occurred when the mosaic plot was included. For example, when Scott was using the mosaic plot in the second task, he said, "just look at the mosaic, it's pretty clear" and provided a succinct explanation.

Interestingly, the mosaic plot seemed to help Klaus as he reasoned through the second task. With the contingency table alone, his additive understanding was apparent and he remained with his initial reasoning relying on the larger numbers. (see Figure 2).

Sport/Instrument	Violin	Saxophone	Total
Soccer	58	36	94
Basketball	13	47	60
Total	71	83	154

"The majority of soccer players play violin. And the majority of basketball players play saxophone."



"So, I don't think those percents mean anything in this situation."

Figure 2. (a) Klaus's work considering greatest numbers in comparison with smaller numbers in a contingency table and (b) computation of conditional probabilities.

Although he computed percentages that he could compare to reason proportionally, he did not recognize their usefulness. However, when he later reasoned with the mosaic plot accompanying the same problem, he clearly used the mosaic plot to compare the two categories and their proportions.

Conclusion

Overall, the mosaic plots appeared to be accessible, appealing, and useful to students. Scott claimed it allowed him to visualize the total as a whole and the percentages better than the table. In addition, all participants agreed mosaic plots were helpful and that having to draw it themselves helped them to understand it. Cici, the youngest participant mentioned it helped her to "memorize it a little more in your head."

Mosaic plots may be a useful representation for students when reasoning about (in)dependence of categorical variables. Future work might consider different aspects of contingency tables, how students understand and work with the constituent components, and how they reason across different representations.

References

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, *52*(3), 215–241. https://doi.org/102431232

Bargagliotti, A., Franklin, C. A., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II)* (Second). Alexandria, VA: American Statistical Association.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, *27*(2), 151–169. https://doi.org/10.2307/749598

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal*, *12*(2), 68–82. https://doi.org/10.26522/brocked.v12i2.38

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, *8*(3), 373–395.

Patton, M. Q. (2005). *Qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Pfannkuch, M., & Budgett, S. (2017). Reasoning from an eikosogram: An exploratory study. *International Journal of Research in Undergraduate Mathematics Education*, *3*(2), 283–310.

Watson, J. M., & Shaughnessy, J. M. (2004). Proportional Reasoning: Lessons from Research in Data and Chance. *Mathematics Teaching in the Middle School*, *10*(2), 104–109.