A CASE STUDY: ZANDER REASONS WITH INCOMPLETE CONTINGENCY TABLES AND MOSAIC PLOTS

Sheri Johnson[1] and Christine Franklin[2]
[1]The Mount Vernon School, Atlanta, Georgia, USA
[2]American Statistical Association (ASA), Athens, Georgia, USA
sjohnson@mountvernonschool.org

Reasoning across multiple representations and incorporating reversibility can enhance and reveal an in-depth view of a student's understanding. In this paper, we explore how a school-age student reasons about the independence of two categorical variables using contingency tables and mosaic plots. This case study reveals some fine-grained reasoning that highlights the differences between mathematical and statistical thinking, shows an inclination to use a part-part or odds approach, and demonstrates how a mosaic plot was used to solve a problem with incomplete contingency. Ultimately, this study supports the use of incomplete contingency tables and mosaic plots for students to reason about statistical independence.

INTRODUCTION

Literacy is an expected schooling result, and statistical literacy requires evaluation and interpretation of data. This includes working with categorical data and summary statistics presented in contingency tables. Statistical concepts such as independence take time to develop, and it is important to introduce them to students early and allow students to grapple with their understanding while reasoning across different contexts and representations (Bargagliotti et al., 2020).

A mosaic plot is a potentially useful tool that assists students in appropriately applying proportional reasoning to determine independence (Pfannkuch & Budgett, 2017). Figure 1 shows a contingency table with the explanatory variable (drug type) oriented horizontally and two associated mosaic plots. Traditional mosaic plots are based on a unit square and show the marginal distribution of the explanatory variable on the horizontal axis (Drug A: 200/480 = 0.417 = 41.7%; Drug B: 247/480 = 0.583 = 58.3%). Mosaic plots also show the conditional distributions within each column for the traditional mosaic plot and within each row for a sideways mosaic plot. These conditional distributions can be compared to conclude whether the categorical variables are likely to be independent or associated.

		Adverse effect?		Row
		Yes	No	Totals
Drug Type?	Drug A	60	140	200
	Drug B	187	93	280
Column Totals		247	233	480

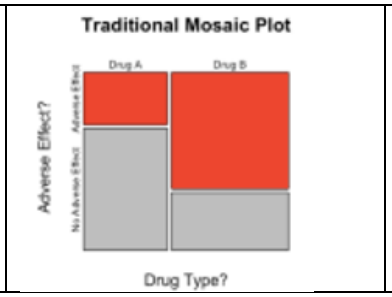

Figure 1. Contingency table, traditional mosaic plot, and sideways mosaic plot

Johnson (2020) used a sideways mosaic plot with the explanatory variable in rows to avoid the Stroop effect, which is when two different representations create cognitive conflict. Pairing these representations allows a natural spatial connection between the joint frequencies in the contingency table and the tiles in the mosaic plot. A benefit of this orientation is that when comparing the length of the bars for each row, they are in the same orientation as a horizontal number line.

BACKGROUND AND MOTIVATION

Many school-age students learn the mantra that association does not imply causation, but do they understand the difference between association (dependence) and independence? From here on independence is referred to as (in)dependence. A contingency table presents the visual relationship

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

between two categorical variables. Whereas the display design looks simplistic, complex relationships exist among constituent components. Understanding these relationships requires statistical reasoning and working with numbers in context, not just numerical fluency. Justifying a conclusion of (in)dependence can be more difficult than identifying the correct conclusion, especially when multiplicative rates beyond halving and doubling are required.

There have been limited studies to investigate students' reasoning about categorical association, including seminal work in mathematics education with upper secondary students in Spain (Batanero et al., 1996). Contradictory contexts, exemplified with smoking and lung disease, in which context-driven preconceptions conflict with numerical responses, revealed that students struggle to let go of preconceptions. This study provided complete contingency tables without additional representations and investigated association.

To learn mathematics and statistics more deeply, researchers encourage reversibility questions, which are characterized by providing results and requiring students to coordinate actions to determine a starting state (Simon et al, 2016). In order to fully understand how students reason with categorical data in contingency tables, it is important to consider how they might choose numbers to make the situation (in)dependent. Reversibility and incomplete contingency tables have been absent from past studies.

Few subsequent studies included or focused on bivariate categorical data using representations beyond contingency tables. Casey, Albert, and Ross (2018) investigated future and practicing teachers' knowledge for teaching the graphing of bivariate categorical data. They used novel curriculum materials that included representations of bar charts with relative frequencies and mosaic plots. Initially, teachers had limited knowledge and only used frequencies to analyze data. After instruction, the teachers were more likely to use relative frequencies and notice inappropriate use of frequencies and labels in student work. Work with younger students (Casey, Hudson, & Ridley, 2018) found that students experienced challenges with using relative frequencies and that younger students reasoned about (in)dependence better using segmented bar charts with percentages than with mosaic plots. However, this preference may primarily be due to familiarity. This study did not use similar representations for students to compare mosaic plots with other representations, and no instruction was provided for a mosaic plot, which is likely an unfamiliar display to students.

METHODS

This case study is part of a larger research effort (Johnson, 2020). The focus of this case study, Zander, was a 14-year-old boy from a suburban middle school in the southeast United States. He was placed in a high level, accelerated mathematics class in grade 8 (13–14 years of age), taking algebra (typically taught in grade 9) and some geometry (typically taught in grade 10). He liked to play video games and read when not in school, and his favorite subject was social studies. He suggested that understanding every concept he learns would help him to like math more. He is a middle child and has an older brother and a younger sister. He sees mathematics as playing an important part in his future.

Zander, like the initial eight participants ages 12–17 in the larger study, was carefully selected, privileging those judged to be metacognitively aware based on past interactions with the researcher. Other relevant criteria considered included students in seventh grade and above who had not taken AP Statistics, with an aim to target students who may have prerequisite skills but have not been exposed to in-depth study of categorical association such as a chi-squared test. Research questions included:

1. In what ways do students reason about (in)dependence of categorical variables when using contingency tables?
2. In what ways do students use mosaic plots to reason about (in)dependence of categorical variables when using contingency tables?

Interview tasks and protocols were developed in consideration of past research, aspects of contingency tables, and a pilot study. For this case study, five interviews were conducted approximately one week apart and began with an initial observational interview to verify Zander's prerequisite ability to reason proportionally, work with simple probabilities, and understand the basic structure of contingency tables because these are necessary skills to reason with contingency tables to determine (in)dependence. Subsequent interviews ranged from probing to assistive to instructional as they progressed to consider Zander's work with complete contingency tables (IV#2—considering relationships from complete contingency tables without mosaic plots and IV#3—considering

relationships with complete contingency tables and mosaic plots) and incomplete contingency tables (IV#4—considering relationships with incomplete contingency tables and without mosaic plots and IV#5—considering relationships with incomplete contingency tables and with mosaic plots). Specific tasks are described in greater detail in the Findings.

The interview tasks increased in their expected difficulty because Zander was not likely to have encountered the terms association or independence previously, and the tasks used questions with the words "less likely," "equally likely," or "more likely." Before reasoning with mosaic plots, Zander was instructed to create a mosaic plot on paper. This entailed dividing a 10 x 10 square grid using a horizontal line based on row marginal frequencies and then further subdividing each partitioned area with a vertical line based on conditional frequencies. Johnson (2020) provides a more in-depth discussion of the methods used in the larger study.

FINDINGS

Zander invoked proportional reasoning and demonstrated other prerequisite skills in this first interview to determine (in)dependence with complete and incomplete contingency tables. When working with complete contingency tables and without mosaic plots, Zander explicitly recognized that when marginal frequencies were equal, he could efficiently reach a conclusion by comparing joint frequencies. When the marginal frequencies were unequal, he primarily used a strategy of comparing conditional relative frequencies to determine (in)dependence. Johnson (2020) includes details about all interview tasks and possible correct solutions.

In the second interview, the country and pop music task (IV#2, Task 5) included unequal row marginal frequencies for two statistically independent variables and asked, "Are middle school students more, less, or equally likely than high school students to listen to country as opposed to pop music?" When Zander worked on this task (see Figure 2), he first used a whole-part comparison and realized his approach was problematic because the resulting numbers were greater than one. He then recognized that a part-whole comparison of the same numbers resulted in a number that was less than one and was easier to use. He used the fact that the conditional relative frequency of preferring country music for middle school students (0.797) was greater than that for high school students (0.796) to conclude that middle school students are "more likely" to listen to country music than high school students.

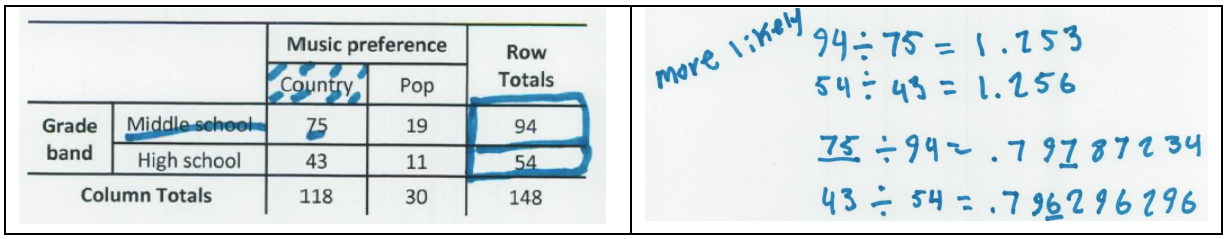

Figure 2. Zander's work for IV#2, Task 5 with statistical independence

Some probing questions about equality prompted Zander to qualify his initial answer of "more likely" with "but almost equal to." Initially, he thought he could create equal ratios by moving just one person. When he tried to move one middle school student from country to pop, however, the ratios revealed a less likely situation ($74/94 = 0.787 < 0.796$). When asked to choose either "more likely" or "equal to," he repeated his initial response of "more likely." This type of problem elicits the difference between mathematical thinking with exact answers and statistical thinking, where context is key and some random chance variation is common. The numbers reveal different rates (probabilities), but they are as close to being equal. Thus, there is a mathematical difference, but not a statistical difference.

When working with mosaic plots, Zander readily noticed the area connection and accurately created a mosaic plot with minimal instruction, only needing clarification of horizontal versus vertical. Throughout IV#3, Zander reasoned across the representations, found mosaic plots useful, and recognized he could compare linear lengths, not just areas of components of the mosaic plot. To first determine (in)dependence, if there was a clear difference in the lengths of corresponding row segments, Zander recognized the more or less likely situation. Alternatively, if the mosaic plot's row

segments appeared to have equal lengths, he looked at the numbers in the contingency table to check to see if the conditional relative frequencies were exactly equal.

When working on the same country and pop music task using a mosaic plot (IV#3, Task 5), Zander noticed that the lengths of the row segments were very close (see Figure 3), and he used the numbers in the contingency tables to check for exact equality. Thus, mathematical thinking remained at the forefront. However, this time he employed an odds approach rather than a risk approach, comparing part-part ratios as opposed to part-whole ratios. Whereas this is a valid approach, it proved to be difficult for him to interpret and he concluded the answer was "less likely," "just by a little." As noted, in the previous interview, he inverted the ratios using a whole-part ratio resulting in a number that was greater than one. This signified an error to him because he understands that a relative frequency or a probability is less than one. Possibly he used that information to wrongly conclude "less likely" rather than more likely. He demonstrated the numerical understanding that the overall odds for all students had to be between the odds for the two groups combined. However, he did not appear to understand the part-part ratios he created.

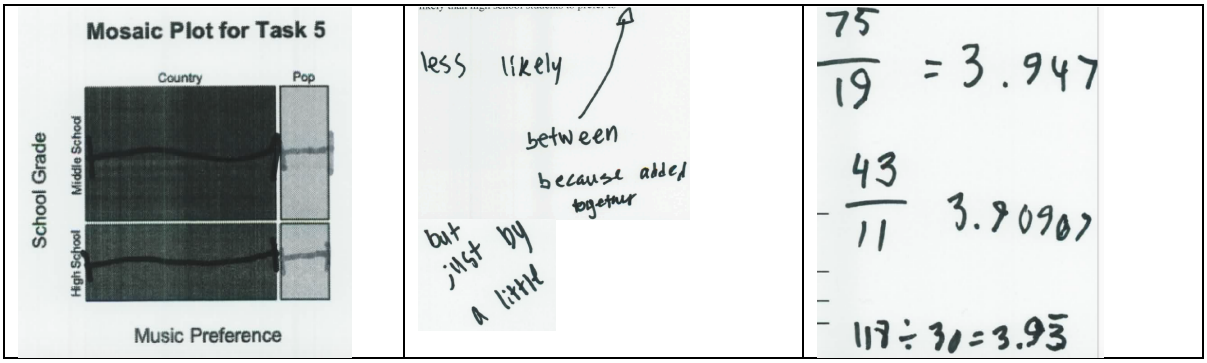

Figure 3. Zander's work for IV#3, Task 5 with statistical independence and a mosaic plot

When working with incomplete contingency tables, Zander was efficient and used zeros for joint frequencies when possible and made use of benchmark fractions (e.g., ½). When given an equally likely situation, Zander used his typical approach of row conditional relative frequencies to determine missing joint frequencies.

However, he struggled with the cereal problem (IV#4, Task 4a), for which all marginal and no joint frequencies were provided, and a condition of equally likely (independent) was specified. This was the one problem with an incomplete contingency table and no mosaic plot that Zander did not complete correctly (see Figure 4), but he later solved it correctly when a mosaic plot was provided. Initially, Zander used a benchmark fraction of ½ to compute joint frequencies, but when he tried to verify the column totals, he recognized that 50% did not work. He acknowledged there was only one set of numbers that worked and stated, "there must be a mathematical way." He tried to determine if the joint frequency was greater than or less than 50%, admittedly using a trial-and-error approach.

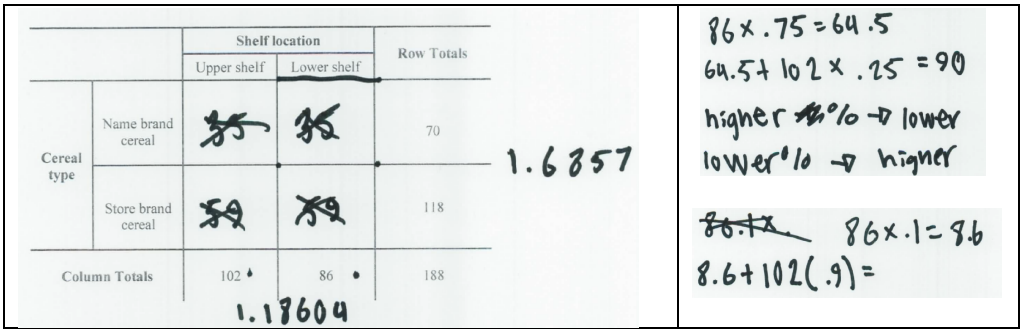

Figure 4. IV#4, Task 4a, cereal task including only marginal frequencies

Zander used the distributive property inappropriately, multiplying complementary percentages by different column marginal frequencies to determine the joint frequencies for the store brand cereal (e.g., 0.25 * 102 = 25.5, 0.75 * 86 = 64.5). Because the sum of these joint frequencies was always less than the marginal frequency for the store brand cereal total of 118, he stopped this strategy. He thought ratios of some sort would help, and he computed the odds (part-part ratios) for each of the row and column marginal frequencies. Then he used one of these ratios and calculated the store brand, lower shelf cell frequency (86 ÷ 1.18604 = 51). He completed the table, recognized this was not an equally likely situation, and asked to stop working on this problem.

When Zander worked on the same cereal problem that included a mosaic plot (IV#5) he correctly coordinated the constituent components (see Figure 5). He calculated the marginal relative frequencies and multiplied them to find the joint relative frequencies for the store brand cereal (e.g., 0.63 * 0.46 = 0.2898). Zander recognized his rounding error after he wrote 28% and used the actual result of 0.2898 to estimate the store brand, lower shelf cereal frequency of 54.4824, which he crossed through because it still had possible rounding errors. He used the memory feature on his calculator to calculate 53.97, which he rounded to 54. He stated, "The mosaic plot gave me the idea first. I was trying to figure out what percentage this was, or the length of this was out of all this side right here".

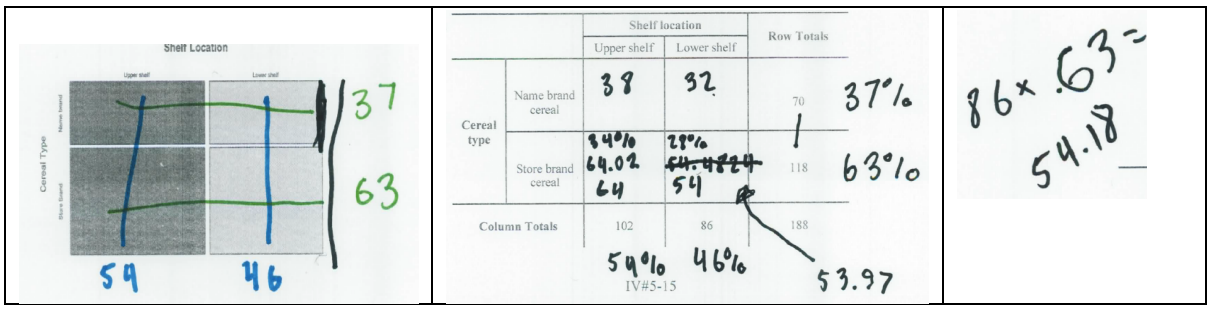

Figure 5. IV#5, Task 4a, cereal task including only marginal frequencies with a mosaic plot

He referred back to his earlier work while working on another problem and describing how the mosaic plot was helpful (IV#5, Task 2a). Zander used a part-part approach for this problem and estimated the number of times the smaller part would fit into the larger part in the same row (see Figure 6).

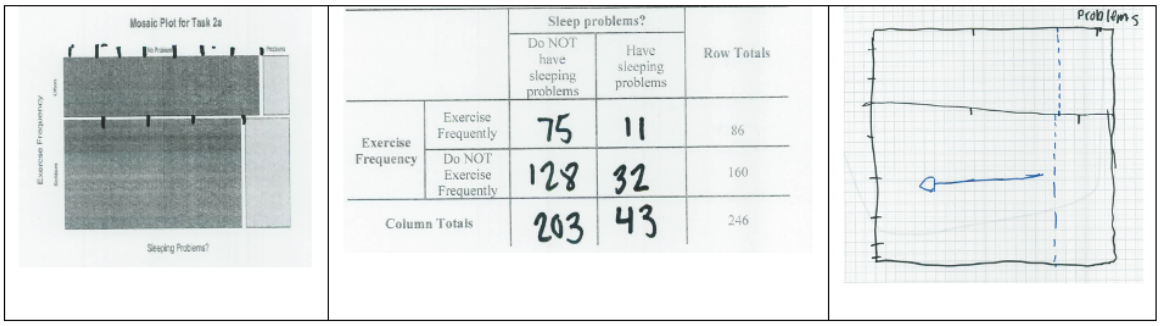

Figure 6. IV#5, Task 2a, incomplete contingency tables and mosaic plot with only row totals

With complete contingency tables, Zander compared the conditional relative frequencies to one another, but when an incomplete contingency table was provided with only marginal values, he needed more information. One approach is to consider that these conditional relative frequencies must also be equal to the marginal relative frequency. Whereas Zander recognized the marginal relative frequencies were important, he did not initially recognize they could be used as a multiplicative operator with the conditional marginal frequency to find the joint frequency. The mosaic plot was what allowed Zander to recognize the marginal relative frequency was important and once he did, he used it as an operator (see Figure 6).

LIMITATIONS AND IMPLICATIONS

Think-aloud clinical interviews provide good insight into student thinking, especially in comparison with written work, but they have limitations of not adequately capturing all thoughts and interrupting the problem-solving strategies being employed. Whereas a single case allows us to consider more fine-grained thoughts, Zander is one individual student with his own experiences. Thus, he alone is not representative of all 14-year-old boys. Nevertheless, this case study does provide some insights into the intricacies of reasoning with contingency tables and mosaic plots.

Working with incomplete contingency tables demonstrates reversibility and a depth of understanding that is not apparent when working with complete contingency tables alone. Zander created and used mosaic plots before looking at the frequencies in the contingency table to determine (in)dependence, tending to both one and two-dimensional measures. A mosaic plot helped him solve a problem with an incomplete contingency table that he was unable to solve without it.

This study supports clarifying the usefulness of different representations such as mosaic plots for younger students who can reason proportionally. What about students who are in the process of developing proportional reasoning? Could mosaic plots, and, more specifically, having students create mosaic plots by hand, be a tool that assists in students' development of proportional reasoning? Mosaic plots have a clear connection with geometry and can be used to reinforce understanding of area, linear length, composite figures, and similarity, thus integrating mathematics and statistics.

Zander naturally gravitated towards a part-part or odds approach but struggled to make sense of and use these ratios as an operator. On the other hand, he seemed to more naturally make sense of a part-whole or risk approach and use these ratios as an operator. When teaching younger students, should there be a clarification between odds and risk? Is the structure of a contingency table a place where they can learn to interpret each ratio and use it appropriately as an operator? Might this be something that helps students to develop multiplicative reasoning as well as better prepare them for future work in statistics?

Incomplete contingency tables and mosaic plots can help to better elicit student reasoning. A better understanding of student reasoning allows us to find the gaps and develop better approaches to create a statistical and data literate society. More research in this area will help to inform student learning, curriculum, professional development, and education of future teachers.

REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II). A framework for statistics and data science education.* American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

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

Casey, S. A., Albert, J., & Ross, A. (2018). Developing knowledge for teaching graphing of bivariate categorical data. *Journal of Statistics Education*, *26*(3), 197–213. https://doi.org/10.1080/10691898.2018.1540915

Casey, S. A., Hudson, R., & Ridley, L. (2018). Students' reasoning about association of categorical variables. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward: Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS 10), Kyoto, Japan*. ISI/IASE. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2E2.pdf?1531364243

Johnson, S. (2020). *Students reasoning about the association of categorical variables using contingency tables and mosaic plots* [Doctoral dissertation, The University of Georgia]. Ex Libris. https://esploro.libs.uga.edu/esploro/outputs/9949365649202959

Pfannkuch, M., & Budgett, S. (2017). Reasoning from an eikosogram: An exploratory study. *International Journal of Research in Undergraduate Mathematics Education, 3*(2), 283–310. https://doi.org/10.1007/s40753-016-0043-0

Simon, M. A., Kara, M., Placa, N., & Hakan, S. (2016). Categorizing and promoting reversibility of mathematical concepts. *Educational Studies in Mathematics, 93,* 137–153. https://doi.org/10.1007/s10649-016-9697-4