Using Student-Generated JMP Output and Analysis to Guide Class Discussion

Elise Lahiere
Montclair State University
June 30, 2021

About the Course

- Introductory statistics course for undergraduate students who were mainly computer science and biology majors
- 38 students enrolled
- Fully online class; one asynchronous meeting and one 75 minute synchronous meeting in Zoom each week
- Flipped classroom; notes were taken and instructional videos were viewed prior to synchronous class meeting
- Each class meeting provided an opportunity for students to interact with the material, either in groups or individually
- Class discussions were centered around student presentations of work completed during class

Overarching Goals

- Engage students during class
  - Active learning, where students are engaging with the material in class, has been shown to have a positive impact on student grades and pass rates (Freeman et al., 2014)
- Build intrinsic motivation to learn and discuss course topics
  - It is important to “motivate how and why data practice is an important topic and focus on the context, impacts, and practicalities of data practice… (which) has the potential to motivate students who are not already inclined to go on to pursue more technical classes and applications” (Donoghue et al., 2021, p. S28)
- Provide insights into statistics, its purpose and practices, focusing on concepts rather than calculations

Connections to GAISE 2016 Recommendations

"Recommendation 1: Teach statistical thinking" (p. 12)

- Students were asked to select variables for analysis, discus findings using context of the data, and think about why the analysis could be beneficial

"Recommendation 2: Focus on conceptual understanding" (p. 16)

- Computations were completed in JMP for the majority of the course, placing more focus on interpretation

"Recommendation 3: Integrate real data with a context and a purpose" (p. 17)

- All data sets used provided contexts, questions provided opportunities to think more deeply about the context of the situation and the motivation for creating outputs and completing analysis, students had freedom to choose which variables to analyze

"Recommendation 4: Foster active learning" (p. 18)

- Students were able to actively engage with the material, discussing statistical ideas with others, and presenting findings using statistical language

"Recommendation 5: Use technology to explore concepts and analyze data" (p. 19)

- Throughout the course JMP was used to complete analysis, other technologies were also used to explore concepts, such as Geogebra and applets

"Recommendation 6: Use assessments to improve and evaluate student learning" (p. 21)

- Student presentations and work in class were used in addition to homework assignments to assess how student understanding was progressing

Week 2: Picturing Distributions with Graphs

- “Visualization (of data) provides one of the most readable and compelling forms of information as well as a highly motivating early experience for beginning students” (Kaplan, 2018, p. 90)
- Topics:
  - Creating and using bar graphs and histograms to visualize and explore features of distributions
  - Connecting features of the graphs to the context of the data
- Topics align with GAISE (2016) goals:
  - “Students should be able to
    - … recognize questions for which the investigative process in statistics would be useful and should be able to answer questions using the investigative process
    - … produce graphical displays and numerical summaries and interpret what graphs do and do not reveal
    - … interpret and draw conclusions from standard output from statistical software packages” (p. 8)

Week 2: Picturing Distributions with Graphs

- Class meeting format:
  - Brief instructor led review of main ideas and concepts that were covered in asynchronous work
  - Group work in breakout rooms with 3 to 4 students
    - Students were provided with a sample data set from JMP
    - Each group selected a quantitative and qualitative variable from the data set, creating a bar graph and histogram in JMP
    - Groups responded to prompts on Google Jamboards
  - Groups presented their findings to the rest of the class

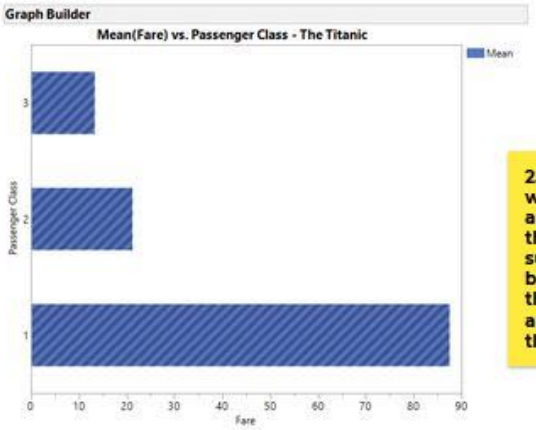
Image source:
https://www.kveller.com/5-things-to-consider-if-youre-signing-up-for-zoom-camp/

Questions for Group Work

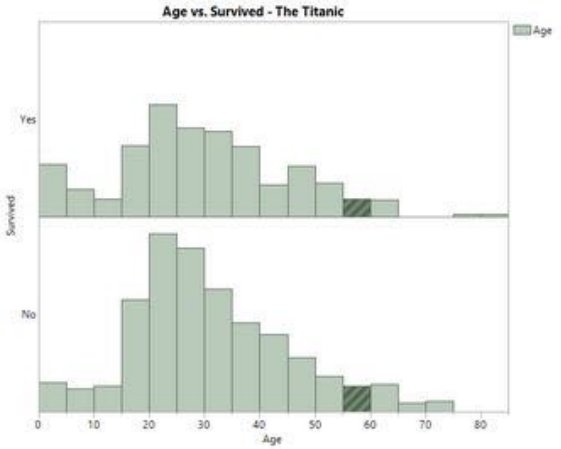
Each group answered the following questions about the graphs they created:

1. What are the individuals in your data set?
2. What variables did you use in each of your graphs?
3. Why did you choose these variables?
4. What are your graphs describing?
5. Who might find these graphs useful?
6. What are two statements that you can make based on each graph? Be sure to use the context of the data for these.
7. What is one question that each graph brings up, but cannot directly answer?

Group Work Responses - Example 1



1. The individuals in this data set are the passengers who were on the Titanic.



2. For the histogram we used the age and compared it to the number of survivors. For the bar graph, we used the passenger class and compared it to the fare they paid.

3.We used these variables because we sensed an interesting connection between them. We feel as though these variables help us understand the disaster better.

4. This graph is describing the passenger classes and how much their fare costs in comparison to each other.

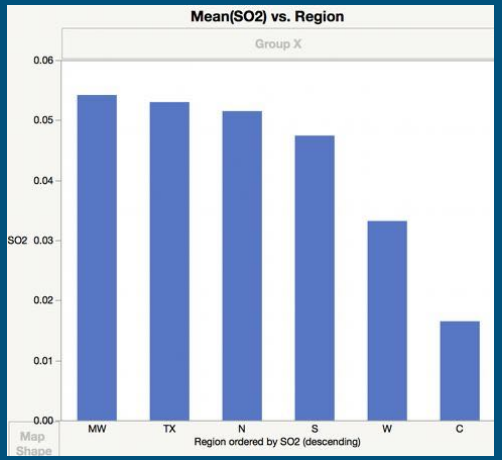
6. Based on this graph, we can safely say that the people in the third class paid significantly less for their tickets compared to the people in the first class. This might reflect on the quality of each passenger class.

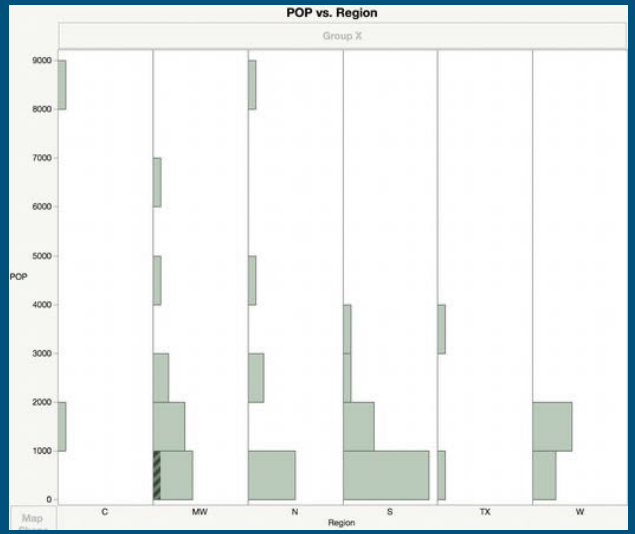
4. This graph is describing the age of the Titanic passengers and whether they survived or not.

6. Both graphs are very similar, so we can safely say that the number of survivors has nothing to do with a specific age group.

5. Lots of different groups of people may find both graphs intriguing. For the bar graph, consumers may want to know the price differential. For the histogram, historians or scientists could find the info useful.

7. One question that both graphs bring up is the number of people in each class, and the number of people that did and did not survive specifically. We know which groups had more over the other, but no specifics.





Group Work Responses - Example 2

- The individuals in our data are cities. The variables we used are regions and SO2 for our bar graph and regions and population for our histogram.
- We chose these variables to show what regions may be at risk for SO2 pollution and what regions are safe. We wanted to see if a higher population directly affects the amount of SO2 emissions because SO2 can be from the burning of fossil fuels
- The first graph shows the midwest having very high levels of SO2 and the north and south have pretty high levels as well. The west and central regions seem to have lower SO2 emissions. In the second graph we actually see that the places with the most SO2 emissions have the lowest population of people. This could be because SO2 emissions come from large power plants and other large facilities that are located away from humans due to the toxic chemical being released. This is something we are assuming but can not be sure of from this data.
- Someone who may find this useful is a doctor that is worried about a patient travelling to a place with high SO2 emissions. Another person who may find this useful is a chemist that wants to test the levels of SO2 in a place with high emissions.
- A statement we could make is the central region has the lowest levels of SO2 but the highest population. The midwest and Texas have the highest levels of SO2 and both have smaller populations.
- A question we have is why do the south and west regions have a smaller population than texas, which is one state? Another question is are there more SO2 emissions in the midwest due to there being more places for power plants?

Topics Discussed in Class Presentations

- Many of the key features of graphs came up naturally in the students' presentations as there were a variety of data sets used and variables selected
- The context of the data was included in all of the student presentations, though not all statements made discussed the context
- Many groups discussed the potential implications of what they found when analyzing their graphs

Image source: https://www.vecteezy.com/vector-art/154406-business-woman-presentation-vector

Observations from Class

- Conversations were richer when compared to a lecture based setting where the instructor provides several predetermined examples to review with the class, this format also provided opportunities for formative assessment
- Many groups of students went beyond the basics of the assignment, creating and comparing multiple graphs as a way to compare different groups
- Students needed to consider the implications of selecting different variables to create their graphs, especially when grouping data using categorical variables

Possible Modifications for In-Person

- Moving to an in-person format
  - Students could work in groups sharing a computer, graphs and responses can still be uploaded to Jamboards for easy presentation at the front of the room
  - If computers are not available during class, students can make graphs prior to class and print a copy, class time can be used to analyze the graphs, presentations can be facilitated with a document camera

Image source: https://stock.adobe.com/search?k=classroom

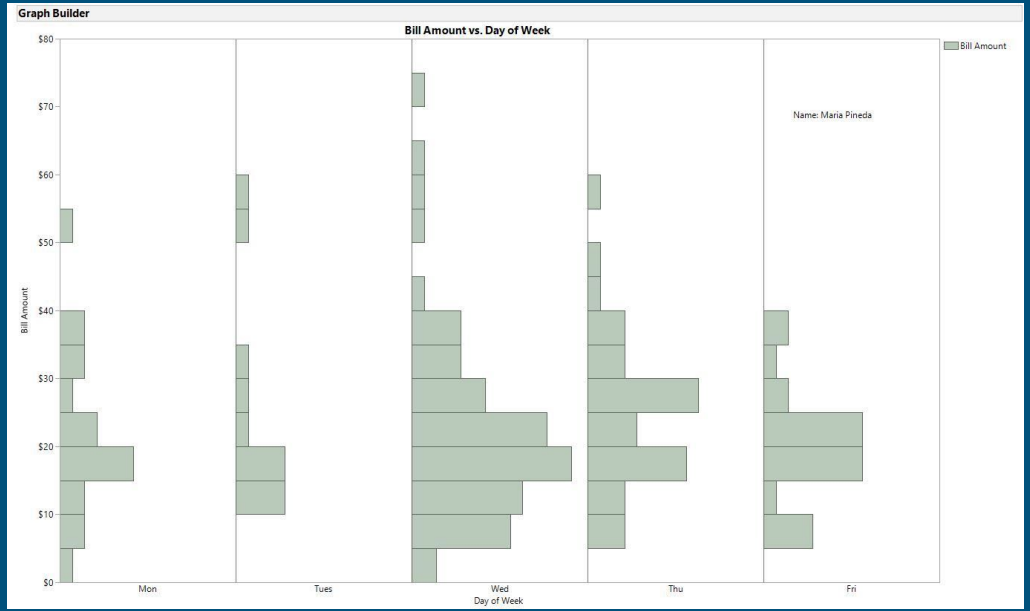
Associated Homework Assignment

After the class meeting the students were each asked to:

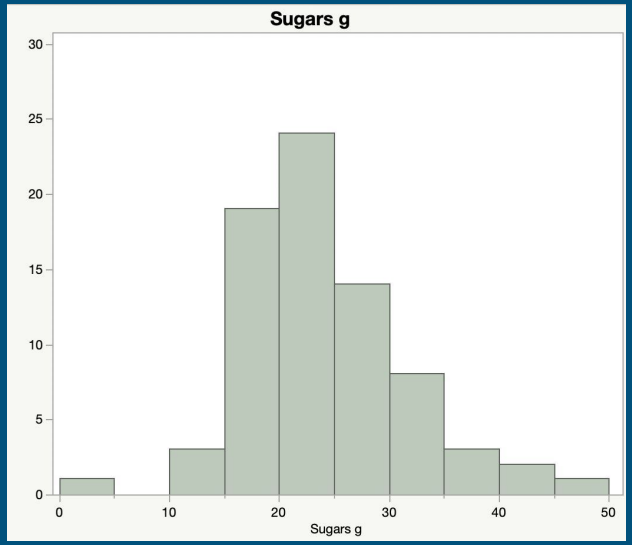
- Create and post a Histogram using a sample data set from JMP
- State the data set and describe the variable(s) used
- Discuss why the graph might be helpful and why they chose to make it
- Discuss the shape, center, and variability of the graph using the context of the data
- State one question that the graph inspires, but cannot answer directly
- Respond to two other posts with a statement that can be made based on another student's graph and a discussion of one interesting result from the graph

Homework Submission - Example 1

- The data set I used is "Restaurant tips".
- I used the day of the week guests came in and the bill amount(how much they spent) as my variables.
- As a restaurant owner or business owners graphs like these help detect what day(s) of the week a guest/customer spends more money. I chose to make it because I wanted to see what day of the week are people more likely to spend money at this restaurant.
- The shape of my graph I'd say looks weird because the data is for 5 different days of the week. The high peaks tell me when there are more people spending a specific amount of money. The low peaks just indicate that there's less people spending money in that range. For example on Thursdays there is more people spending $30 than on Wednesdays, but on Wednesdays there is more people spending $20 than on Thursdays.
- The graph can inspire on what days does this restaurant make more revenue. It does not specifically answer it but the guests spending habits does inspire this question. Another question that can be asked is on what times of each day are these bill amounts the highest? The morning?Afternoon? Etc.



Homework Submission - Example 2



- I used the Candy Bars data set. In the histogram, I used sugars (in grams) as my variable.
- Consumers or manufacturers of Candy bars can use this graph to help give them an idea of the amount of sugar in grams found in candy bars so that they can make healthier choices. I chose to make this graph because I like to know the amount of sugar I consume when eating Candy bars.
- The distribution of the graph seems to be bell-shaped and almost symmetrical. This shape showed me that the majority of Candy Bars have between 20-25 grams of sugar and that fewer candy bars tend to contain sugar over that amount. The data set had a total of 75 observations which ranged from 1 to 47 grams of sugar. By arranging them in order of size I discovered that 23 was the midpoint of the data's distribution, making 23 the center of the distribution. Since the graph is bell-shaped, the center, in this case, is a good analysis of what's happening, which is that most Candy Bars in the data set have approximately 23 grams of sugar in them. The amount of sugar in grams ranges from 1 to 47 which shows the variability.
- Finally, one question that this graph inspires but can't answer would be why do most Candy bars have between 20 to 25 grams of sugar? Is it because most people prefer Candy Bars with sugar in that range or does it have something to do with companies wanting their candy bars to be a little bit more nutritious?

Observations from Homework

- Students were able to individually create and analyze graphs, considering the context of the situation based on the data set and variable(s) they selected
- Most students provided thoughtful questions that the graphs raised, but could not directly answer, extending the thought process to potential further analysis
- A good number of students went beyond the basics of the assignment, creating and comparing multiple histograms as a way to compare different groups

Image source: https://www.dreamstime.com/photos-images/word-homework.html

Student Feedback

Students also found that this format worked well for them. A sample of responses from an anonymous survey given in the third week of class (a week after the activity presented here) are provided below:

- “I find it helpful to discuss what we're doing in the class and working together to figure out what we should do for our graphs and how to compare them.”
- “I've taken a statistics course before (more math based), so using technology that allows us to manipulate the data is very interesting. “
- “I really like how my group works well each other and they often help me if I am having trouble understanding a topic.”
- “I really like it because I get to learn about graphs which I never thought I could do it.”
- “I like how we work in breakout rooms, it keeps me engaged in class.”
- I like the “Analytic nature / uncovering truth (behind numbers) / making sense of things (from data)

(Some students did mention that there were individuals in groups who did not contribute as they should, this is difficult to monitor in an online course and may be easier to mediate in an in-person class)

References

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Holmes, G., Belleman, P., Witmer, J., & Wood, B. (2016), “Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016”

Donoghue, T., Voytek, B., & Ellis, S. (2021) Teaching Creative and Practical Data Science at Scale, *Journal of Statistics and Data Science Education*, 29:sup1, S27-S39.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014), “Active Learning Increases Student Performance in Science, Engineering, and Mathematics,” *Proceedings of the National Academy of Sciences of the United States of America*, 111, 8410–8415.

Kaplan, D. (2018) Teaching Stats for Data Science, *The American Statistician*, 72:1, 89-96.