Building Informal Statistical Inference about Correlation

Elise Lahiere, Katie Monahan, Youngjun Kim
Montclair State University
July 1, 2021

Background

Correlation is particularly important in a society where a large number of information is available because it enables students to explore the strength and direction of relationships between two variables (GAISE College Report ASA Revision Committee, 2016; Goodwin & Leech, 2006).

However, many students “hold misconceptions about correlation, even after receiving instruction” (Liu et al., 2009, p.792).

Misconceptions (Liu et al., 2009)

- The association of a positive correlation coefficient with stronger relationships than negative ones
- Correlation implying causation
- Negative coefficients indicating that no correlation is present

Important but not widely covered concept (Goodwin & Leech, 2006)

- The effect that the presence of outliers has on the correlation coefficient

Goals and Research Question

Goals:

- To create a series of activities that will allow students to learn about correlation in a dynamic setting rather than memorizing the behavior of the correlation coefficient
- To enable students to discuss their ideas in a community of discourse

Research Question:

- What statistical ideas are present in student's discourse as they learn about correlation using discovery methods?

Design

Activities

- Series of 5 tasks within a GeoGebra book: https://www.geogebra.org/m/d2epbt2s
- Focus on building understanding of correlation without previous instruction

Data Collection

- Introductory statistics course for Biology majors
- 50 minutes class session in Zoom
- 7 students took part, split into 3 groups
- Data was collected through
  - Small group conversations
  - Written responses in GeoGebra

Image Source:
https://www.vecteezy.com/vector-art/2497013-online-class-study-from-home-via-teleconference

A Framework for Informal Statistical Inference

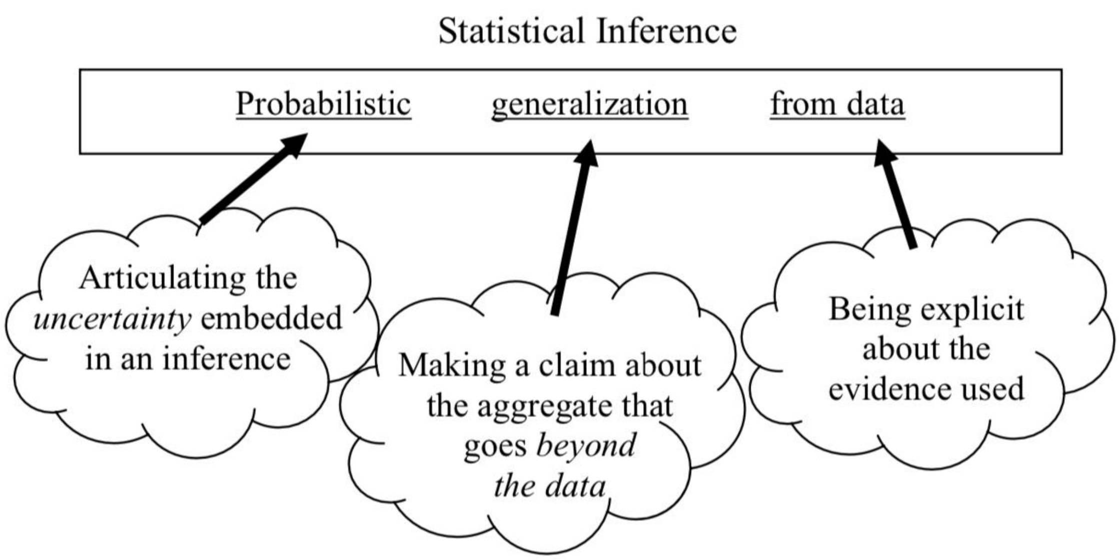


Figure 1: A framework for thinking about statistical inference

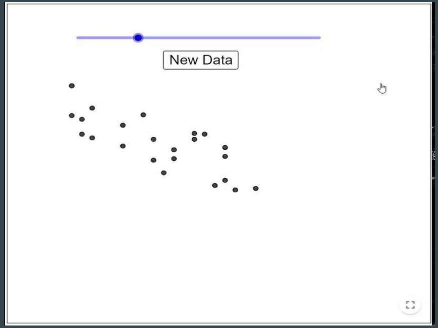
Makar & Rubin, 2009

Building a Model of the Correlation Coefficient (Tasks 1 & 2)

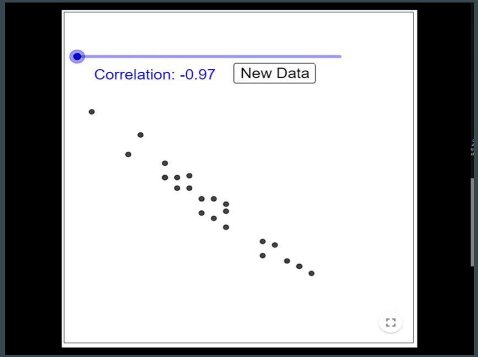
Goal: Build a model for how the correlation coefficient works

Students were asked to:

- Explore dynamic scatterplots
- Describe what they notice
- Build a set of categories to describe the patterns
- Connect the categories to the correlation coefficient



https://www.geogebra.org/m/pjguwyhd



https://www.geogebra.org/m/tvujftgh

Building a Model of the Correlation Coefficient (Tasks 1 & 2)

Informal Statistical Inference Observed: Generalizing beyond the Data

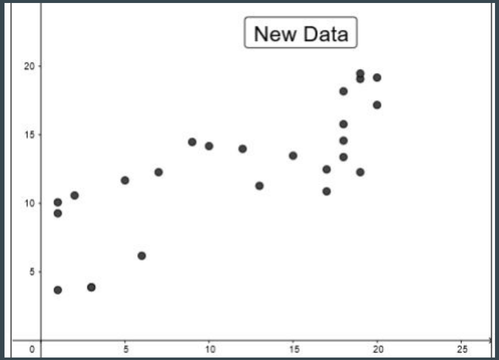
- Generalization beyond the data provides the foundational inferential lens to move from describing the given data and shift towards the target of the inference.
- All students were able to move from describing the given set of scatterplots to describing a general framework for how the correlation coefficient behaves

Researcher: What do you notice about the graphs?

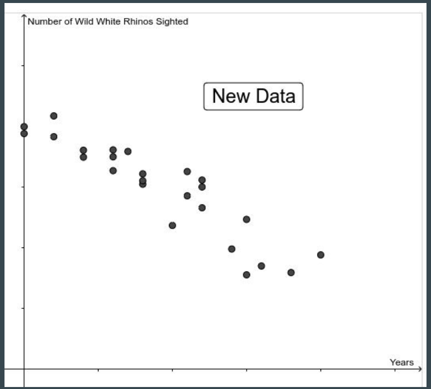
Megan: I said, that if you put the slider closer to either end, the data becomes closer together and formed a straighter line. And if you put the slider towards the middle of the data becomes more spread out and it deviates from the straight line.

Pat: In the first test, I noticed that when I moved the slider, it changes between like positive and negative correlation.

Applying the Model to Scatterplots (Tasks 3 & 4)



https://www.geogebra.org/m/xenkhr3z



https://www.geogebra.org/m/ueqaadhz

Goal: Formative assessment to ensure students build an appropriate model of correlation, providing opportunity to extend that understanding by applying model to contexts

Students were asked to:

- Describe patterns in static scatterplots
- Identify an estimated correlation coefficient
- Interpret graphs that provided context

Applying the Model to Scatterplots (Tasks 3 & 4)

Informal Statistical Inference Observed: Using probabilistic language

- Six of the seven students made statements that provided evidence of statistical uncertainty through probabilistic language
- While this activity also supported statistical inferences such as using data as evidence, the discussions showed the clearest incidences of probabilistic language.

"as the years went, kept going on, so I guess like over maybe 5 years, the number of rhinos could have been at 100 and then they went down to like 50, and, those aren't really good numbers, but I guess as the years go continuing the number of rhinos just go down. I don't know if they're at the same rate as they go down, probably not, but that's what overall thing with the graph is"

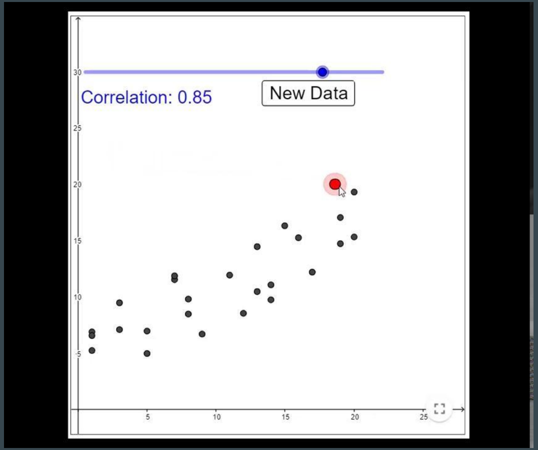
"I said that as the number of rhinos sighted increased the number of years increased as well, but … I wasn't sure because it's not super strong in the, I guess the spread"

Extending the Discussion with Outliers (Task 5)

Goal: Provide extension that allows students to observe the impact that outliers can have

Students were asked to:

- Explore effect of a movable point on the correlation coefficient



https://www.geogebra.org/m/swycuzyk

Extending the Discussion with Outliers (Task 5)

Informal Statistical Inference Observed: Generalizations beyond the data & use data as evidence

- All of the groups were able to describe the effect that outliers would have on scatterplots with different directions and strengths
- Students worked to build a model for how individual points can influence the correlation coefficient, and naturally began distinguishing between points that are close to the overall pattern of the graph and those that are far away.

Megan
(describing a scenario for a scatterplot with a strong negative trend)

"as you move the red dot further away from the other data points, the correlation becomes weaker. And when the red dot is close to or in line with other data points the correlation becomes stronger."

Conclusions

Research Question: What statistical ideas are present in student's discourse as they learn about correlation using discovery methods?

- Using the sequence of tasks in GeoGebra, all groups were able to discover how the correlation coefficient behaves, and how an outlier affects the correlation coefficient

Analysis with Framework:

- All students made comments that could be categorized as Generalizations Beyond the Data and as Data as Evidence
- All but one student made comments that would be classified as Probabilistic Language

References

Arbain, N., & Shukor, N. A. (2015). The effects of GeoGebra on students achievement. *Procedia-Social and Behavioral Sciences, 172*, 208-214.

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM, 44*(7), 913-925.

Engel, J., & Sedlmeier, P. (2011). Correlation and regression in the training of teachers. *In Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 247-258). Springer, Dordrecht.

Frischemeier, D., Biehler, R., & Engel, J. (2016, July). Competencies and dispositions for exploring micro data with digital tools. In *Paper in preparation for IASE Roundtable Conference*.

GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," http://www.amstat.org/education/gaise.

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics, 31*(3), 72-77.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education, 74*(3), 251-266.

Gordon, S. P., & Gordon, F. S. (2018). Visualizing and understanding regression and correlation using dynamic software. *International Journal for Technology in Mathematics Education, 25*(2), 23-32.

Gravemeijer, K., & Bakker, A. (2006, July). Design research and design heuristics in statistics education. In *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-6).

References (continued)

Hall, J., & Chamblee, G. (2013). Teaching algebra and geometry with GeoGebra: Preparing pre-service teachers for middle grades/secondary mathematics classrooms. *Computers in the Schools, 30*(1-2), 12-29.

Khairiree, K., & Kurusatian, P. (2009). Enhancing students' understanding statistics with TinkerPlots: problem-based learning approach. In *Proceedings of the 14th annual ATCM conference*.

Lee, H., & Hollebrands, K. (2008). Preparing to teach mathematics with technology: An integrated approach to developing technological pedagogical content knowledge. *Contemporary Issues in Technology and Teacher Education, 8*(4), 326-341.

Liu, T.-C., Lin, Y.-C., & Tsai, C.-C. (2009). Identifying senior high school students' misconceptions about statistical correlation, and their possible causes: An exploratory study using concept mapping with interviews. *International Journal of Science and Mathematics Education, 7*(4), 791-820.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

Noll, J., & Kirin, D. (2017). TinkerPlots model construction approaches for comparing two groups: student perspectives. *Statistics Education Research Journal, 16*(2).

Prodromou, T. (2014). GeoGebra in teaching and learning introductory statistics. *Electronic Journal of Mathematics & Technology, 8*(5), 363–376.

Sánchez, E., & Izunza, S. (2006). Meanings' construction about sampling distributions in a dynamic statistics environment. In Proceedings of the Seventh International Conference on Teaching Statistics. Salvador (Bahia), Brazil: *International Association for Statistical Education*. CD ROM.