EXPLORING INFORMAL INFERENTIAL REASONING THROUGH DATA GAMES

Kosoom Kreetong
University of Massachusetts - Amherst, United States
kkreeton@educ.umass.edu

This study explores the use of the computer game Ship Odyssey to facilitate learning of ideas underlying students' informal inferential reasoning as they solve parameter estimation problems. I observed seventh graders as they played the game. Players send rats to locate sunken treasure. The rats return with 'noisy' readings. Players determine the treasure location using these data. Thus the game requires parameter estimation in a repeated measurement context, a context that various researches (Konold & Pollatsek, 2002; Lehrer & Kim, 2009) have claimed to be particularly conducive for learning to conceive of data as signal and noise. This conception is precursory to understanding that larger samples give better estimates of a population or process signal.

INTRODUCTION

Informal inferential reasoning is commonly referred to as a process of using data in making arguments and drawing conclusions about some broader context where the conclusions are understood to come with some level of uncertainty (e.g., Zieffler, Garfield, Delmas & Reading, 2008; Makar & Rubin, 2009; Makar, Bakker & Ben-Zvi, 2011; Rossman, 2008). Informal inferential reasoning incorporates several ideas in statistics such as properties of aggregates (data as a combination of signal and noise, types of variability), the effect of sample size on accuracy of an estimate of a population or process signal (Rubin, Hammerman, & Konold, 2006; Pfannkuch, 2005), and uncertainty (Makar & Rubin, 2009; Rubin, Hammerman, & Konold, 2006).

In the recent years researchers have been studying the development of informal inferential reasoning and have made some progress in answering questions on how it can be characterized and how it can be promoted in students and adults. With technology providing new affordances to students' learning experiences in statistics, it is important to continue to explore how technology tools, problem contexts, and instructional sequences can facilitate the development of statistical ideas underlying informal inference.

This study explores the progression of thinking as students engage in parameter estimation tasks in a repeated measurement context in the game *Ship Odyssey*. The study was motivated by the results from earlier research on students' notion of average and the methods they choose to summarize data. Several studies have reported students' tendency to summarize data by using the mode, the midrange (Russell, Schiefter, & Bastable, 2002), and informal methods such as "modal clump"—a range of data in the middle of a distribution (Konold, Robin, Khalil, Pollatsek, Well, Wing, & Mayr, 2002). Building on these results, Konold and Higgins (2003) suggested what young students have in mind about their 'ideal' average—that the average should be "*an actual value in the data set*" that is also "*the most frequently occurring value,*" "*positioned midway between the two extremes,*" and "*not too far from all the other cases*" (p. 203). They also pointed out that the belief that an average needs to be an actual data value could account for students' dismissal of mean as a useful average. Ship Odyssey and the associated activities were designed to help students come to understand that: 1) mean is a useful estimator from which they can make an inference about the value of an unknown parameter; 2) sample size matters—the larger the sample, the better estimator the mean is for making an inference about the value of an unknown parameter. We are exploring ways the game and activities can support the development of these ideas, which are fundamental to informal inference.

The major purpose of the study was to improve the game and the classroom activities written around it. My primary research question was: How does students' informal inferential reasoning develop as students interact with the Ship Odyssey game? This paper focuses on two seventh graders: Ellen (El) and Natalie (Nat).

BACKGROUND

Computer game play generates rich and interesting data, which usually disappear when the game ends. In the Data Games Project, we have created an environment where games are embedded in a dynamic data analysis tool, and data generated from game play remains accessible after the game ends. This allows players to make use of data to improve their game play.

Ship Odyssey is one of the games developed as a part of the Data Games project. In Ship Odyssey, players send rats down to locate sunken treasure. Rats swim back with ‘noisy’ readings. Players estimate the treasure location using these data.

Figure 1 is a screenshot from Ship Odyssey. To obtain measurements of the treasure location, a player presses the “Send Rats” button shown on the upper left. The rat readings appear as dots in the graph on the right. Several tools are available from the gear menu on the upper right corner of the graph, including displaying means and medians on the graph. On the upper right of the game screen is a hook control panel where a player positions and drops the hook. Information about the interval that the hook covers (a total of 4 units) is also provided there.

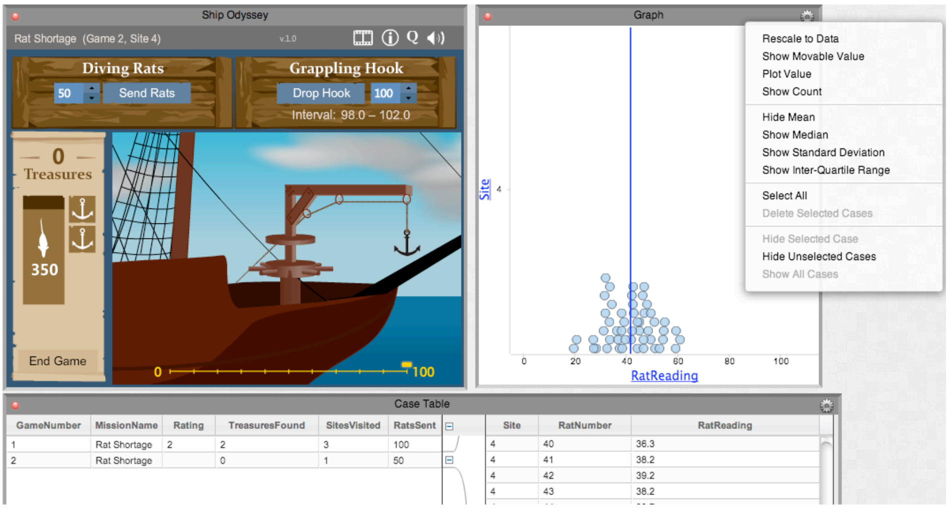

Figure 1: A screenshot from the game Ship Odyssey with the graph menu options visible

One learning objective of the game is that one can make use of ‘noisy’ data to make estimates of unknown population/process signals. Another objective concerns the idea of uncertainty: to understand that failure to find a single treasure does not mean that a strategy is wrong, that in fact even the best strategy will fail sometimes. The game has six missions each targeting slightly different objectives. In each mission players are given a certain number of rats to capture as many treasures as possible. The mission ends when players have missed with the hook twice or they run out of rats. The hook is four units wide except in the mission *Small Hook*. Missions included in this study are *Hundreds O' Rats* and *Rat Shortage*.

Hundreds O' Rats aims to motivate the use of the mean and/or median as an estimate for a treasure location. The mission starts with 800 rats. Players can only send 100 rats at a time. Players will capture the treasure about 96% of the time if they use the mean or median of 100 rat readings whereas other informal methods will yield far lower success rates (about 25% for the apparent mode and about 50% for both the midrange the center of modal clumps).

Rat Shortage aims to motivate using samples that are large enough so that one would not miss too often. The mission starts with 400 rats. Players can send any number of rats at a time. Players will capture the treasure about 80% of the time if they used the mean or median of 40 rat readings.

METHOD

Context of study

Two classes of seventh graders (N=26) participated in two, one-hour long activities designed around the missions *Hundreds O' Rats* and *Rat Shortage*. The ideas we hope to develop are 1) with samples sizes of 100, the mean and median are very good estimates for the treasure location, and 2) the larger the sample, the better estimator the mean/median is for the treasure location.

Activities Design

In the first activity, students played *Hundreds O' Rats*. They sent down 100 rats at a time to locate a treasure. In this activity, students first located the treasure using their own methods. This was followed by a class discussion, which included listing all of the different methods that students had used to locate the treasure. Common methods included the high point of the graph and modal clumps. Mean and median were introduced by the instructor if they were not suggested or used by one of the students. Students then were assigned to play the game again using one of the listed methods to test how well it did. The combined class results showed that mean and median were much more effective estimators of the treasure location than the informal methods students used.

In the second activity, students played *Rat Shortage*. The difference from the first mission was that they could send down as few rats as they liked. They soon found that even though they used mean and median, sample size mattered; if they sent down a small number of rats (e.g., 5-20), they often missed the treasure even though they used the mean or median. This was followed by a whole-class discussion, which included a demonstration of the mean and median stabilizing as a sample grows larger in size.

Data Collection

I used screen-recording software to capture game play and conversation from 7 students (3 males and 4 females). The students worked in pairs so that discussion about their methods and reasons occurred rather naturally as they negotiated how many rats to send down and where to drop the hook. Three of these students participated in group interviews, which took place during the class periods as students worked with the computers. I acted as a participant observer, interacting with the pairs to probe for more information about their thinking. The focus of the interview was on the decisions they made during game play and the reasoning behind their decisions.

Data Analysis

From the video transcript, I develop a *descriptive analysis*, which provides a statement of what the participants intended to express directly (Konold & Well, 1981). From the descriptive analysis, I develop an *interpretive analysis* where I made inferences about the participants' reasoning processes (Konold & Well, 1981). The preliminary findings presented below are based primarily on the group interviews with El and Nat.

PRELIMANARY FINDINGS

Background on El and Nat

El and Nat were selected to be part of the group interviews by their teacher based on the fact that they tended to be more outspoken than others and also worked well together. During the game play they discussed decisions together, although it was El who had the mouse and who seemed to have the final say.

El and Nat playing Hundreds O' Rats

On their first encounter with the Hundreds O' Rats, El suggested

El: *They [rat readings] are mostly right here, don't you think?*

They were both trying to eyeball the location of the highest point in the graph and agreed to drop the hook at its location, 21.

Researcher: *Why did you choose 21? I'm just curious.*

El: *Because they [rat readings] are mostly around 21.*

When this failed to capture the treasure, they looked to the next highest point. Using it, they failed again. They seemed frustrated that their method was not working. El hit the table with her mouse and muttered under her breath "How is that..?" They did not appear to know that the high point of the graph is highly dependent on scale. That is, if they were to resize the graph, they would see that the high point changes location because of the way the program packs points together. Thus this high point is usually not the mode. Nevertheless, they continued using the highest peak of the graph of 100 rat readings to locate the treasure, and if that failed, proceed to the next highest peak. In the second game they were more successful. When asked whether they thought their method (which they later called "highest peak") had worked well so far, they said that it depended on the shape of the distribution.

El: *It depends, like, if there's a lot that go straight down and… or if they're all clumped up. Because when they're clumped up it, like, spreads more. So if they are all, like, straight up and down, it's easier to tell.*

It appeared to El that the highest peak method worked well when the distribution had a predominant peak ("straight up and down" as described by El) so it was easier to see the highest peak. But this method did not work as well when the distribution was more spread out with no definite peak ("clumped up" as described by El).

Following the class discussion of methods, El and Nat were assigned to test the median, which they could add to the graph using the menu options. They used a rounded integer value of median (which is displayed to the nearest tenth) and found that the median worked better than their method of using the highest peak. When asked what they thought about median, they replied:

El: *(nods) It's been working.*

Nat: *better than the first thing we did.*

El: *Yeah, better than the highest peak one.*

El and Nat had been using a rounded integer value of median and captured a treasure 13 times in a row before their first missed hook drop. On their 14th attempt, the median of 100 rat readings was 35.22. They dropped the hook at 36 and missed. They adjusted the hook to 35, dropped the hook and got the treasure. Their actions suggest that the information they obtained from the missed hook was that the treasure was not at 36 when in fact they potentially now knew that it was not in the interval 34 – 38. If they had known this, they presumably would have dropped the hook to cover the interval 30 – 34. Thus it seems their strategy was to use the integer value of the median, and if that failed to use the integer immediately above or below. This approach is akin to what they did with the highest peak method—if the first highest peak failed, use another.

El and Nat playing Rat Shortage

During the second activity, El and Nat played nine games of *Rat Shortage. In this mission* they could control how many rats they sent down. They initially chose to send down 6 rats and positioned the hook at 95—a rounded integer value of the median of those 6 readings (94.91). When this failed, El concluded:

El: *There's not enough data to do the median. Maybe there's not enough data.*

In this statement it is not clear whether she believes the median cannot be computed with six values, or whether it is not reliable. They sent nine more rats. The new median was 95.10. El pointed out that the median was still 95 and 96 was still in the interval. Notice that they were now aware of the interval that the hook covered. Seeing that the value of the mean was 97.37 they agreed to drop the hook at 98 and missed. In the next game, they sent 15 rats and got the median. When the hook missed, they began exploring other options in the graph menu. They first added to the graph the standard deviation (SD) and then the inter-quartile range (IQR) (Fig. 2).

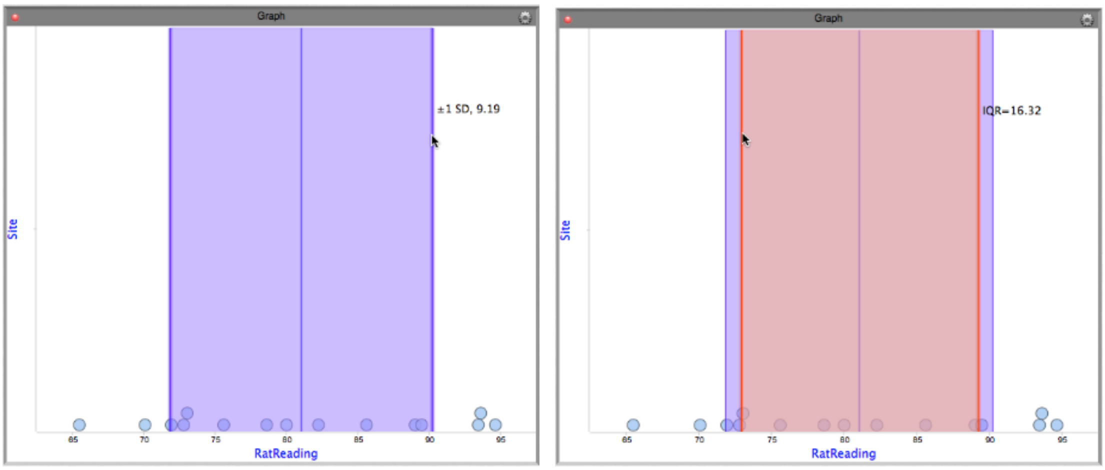

Figure 2: Graphs showing SD (left) and both SD and IQR (right)

El turned to Nat and suggested that they try dropping the hook at 17.

Researcher: *17 based on what? Based on that (pointing at IQR value displayed on graph)?*

El: *Yeah.*

Researcher: *Do you know what it means?*

El: *No*.

The hook came up empty. A feedback screen showed an instant replay of where the first and the second hook drops were as well as the treasure location. El noted that the treasure location was 75. They went back to the graph to inspect the values of IQR and SD and then turned them off and turned the mean back on.

Note that when they positioned the hook at 17, based on the IQR, they placed the hook far to the left of the lowest rat reading at 66. My interpretation is that at this point they were no longer attending to the data in the graph; rather they were looking for THE number to use to position the hook, believing that somewhere in the graph menu was an option that gave them that number. If it was not the mean or the median, then it must be something else, maybe the IQR. They saw that neither the SD nor IQR gave the answer they were looking for (75), but the closest thing was the mean (81.05). In the next games, they used the mean with the number of rats doubled to around 30 and got the treasure. As the games progressed, they noted that more rats work better.

El: *It's easier to find it [the treasure] when you have more rats… It's harder to find it when you don't have a lot of rats 'cause the spots are varied.*

It was not long before they missed again. They switched to median on subsequent drops. About midway through the activity, it seemed they were about to settle on using the median of 40 rat readings. With that strategy, they had captured six treasures before they missed, and the miss seemed to surprise El.

El: *Doh! I thought we were doing well.*

After that one 'surprise' failure, they returned to using the median while varying the number of rats. Typically, when they failed to capture the treasure, they increased the number of rats at the next site whereas when they succeeded, El decreased the number of rats at the next site despite Nat's protestations. Throughout the activity, El and Nat debated about how many rats to send down. Although they both had at some point expressed that "the more rats, the better," Nat generally wanted to send more rats so that they would not miss whereas El wanted to send fewer so that they could potentially get more treasures.

They both seemed to understand that the chance of getting the treasure increased with more rats although they did not express this in probabilistic terms.

El: *You would have to be really, really lucky if you send down one rat and get the treasure.*

Nat: *[In reaction to their failure with the median of 34 rat readings] We got it at 30 [rats]. I guess we weren't so lucky this time.*

The activity came to an end without them settling on any particular number of rats. The number of rats they sent before dropping the hook ranged from 6 to 53. The most consistent aspects of their approach were 1) the use of median and 2) abandoning a strategy after one failure.

DISCUSSION AND CONCLUSIONS

The above analysis sheds some light on the progression of El and Nat's thinking as they engaged with parameter estimation tasks in Ship Odyssey. It also illustrates the challenges we face in getting students to use formal averages, such as means and medians, in meaningful ways.

The informal method of using the high point of the graph readings was not effective at locating the treasure, but at least they were using the data in the graph in a sensible way. I did not ask them why they used the high point of the graph, but we might presume that it is based on the belief that the most common rat reading has the highest chance of being the correct location. Once they were introduced to mean and median and saw that they were more effective at locating the treasure than the method they had been using, they appeared to abandon their informal knowledge and switch to mean and median. But in the first activity, there is no development of the reason that these work—we expect students to begin using them initially because of their proven success. That they used in one instance the IQR to position the hook at a location that was far away from any of the rat readings is a good indication of just how meaningless the task had become for them at that point. We do not yet know how common this response might be, but it at least suggests the need for a revision of the activity. This result is consistent with earlier reports about students' intuitive averages (e.g., Konold et. al., 2002; Russell et al., 2002).

Their typical reaction to a missed hook drop and their comments along the lines of "the more rats, the better" suggests that they have some intuitive understanding of the effect of sample size. Their remarks about getting lucky reflect some intuitive understanding of chance. However, their dismissal of a 'good' strategy (the median of 40 rat readings) after one failure suggests that they may believe that an effective strategy would be one that never failed. We did not manage in these two activities to get them thinking probabilistically about inferences from random samples, that a good estimating strategy will still fail sometimes and a poor strategy will sometimes succeed.

Nevertheless, students were deeply engaged. They were excited to test different methods of locating the treasure and intrigued by the results. This was likely their experience of using the mean or median to actually achieve an objective. And they seemed captivated by the demonstration that the mean and median stabilize as the sample size grows larger.

It appears that students already have bits and pieces of ideas foundational to informal inference, but these ideas are somewhat disconnected. We are continuing our classroom-based research and revising the activities to facilitate their development.

ACKNOWLEDGMENT

I wish to thank all the students who participated in this study and their teacher. I would also like to thank Dr. Cliff Konold for his generous feedback on the research instruments and the manuscript. The study was funded by National Science Foundation, award DRL-0918653. The views expressed here do not necessarily reflect the views of the Foundation.

REFERENCES

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Konold, C., Robin, A. , Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). *Students' use of modal clumps to summarize data*. Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Konold, C., Well, A. (1981). *Analysis and reporting of interview data*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Leher, R., & Kim, M. (2009). Structuring variability by negotiating its measure. *Mathematics Eductation Research Journal*, *21*(2), 116-133.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*, 152-173.

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

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. A. Jones (Ed.), *Exploring Probability in School: Challenges for Teaching and Learning* (pp. 276-294). New York: Springer.

Rossman, A. J. (2008). Reasoning about informal statistical inference: one statistician's view. *Statistical Education Research Journal*, *7*(2), 5-19.

Rubin, A., Hammerman, J. K., & Konold, C. (2006). Explore informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds), *Proceedings of the Seventh International Conference on Teaching Statistics*. Salvador, Bahia, Brazil: International Association for Statistical Education. Retrieved from http://iase-web.org.

Russell, S. J., Schiefter, D., & Bastable, V. (2002). *Developing mathematical ideas: Working with data*. Parsippany, NJ: Dale Seymour Publications.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistical Education Research Journal*, *7*(2), 40-58.