EXPLORING INFORMAL INFERENCE WITH INTERACTIVE VISUALIZATION SOFTWARE

Andee Rubin and James K. L. Hammerman
TERC, United States
Cliff Konold
University of Massachusetts - Amherst, United States
andee_rubin@terc.edu

Most statistics educators would agree that statistical inference is both the central objective of statistical reasoning and one of the most difficult ideas for students to understand. In traditional approaches, statistical inference is introduced as a quantitative problem, usually of figuring out the probability of obtaining an observed result on the assumption that the null hypothesis is true. In this article, we lay out an alternative approach towards teaching statistical inference that we are calling "informal inference." We begin by describing informal inference and then illustrate ways we have been trying to develop the component ideas of informal inference in a recent data analysis seminar with teachers; our particular emphasis in this article is on the ways in which teachers used TinkerPlots*, a statistical visualization tool. After describing teachers' approaches to an inferential task, we offer some preliminary hypotheses about the conceptual issues that arose for them.*

INTRODUCTION

Most statistics educators would agree that statistical inference is both the central objective of statistical reasoning and one of the most difficult ideas for students to understand. Recent research of our own (e.g., Rubin and Hammerman, 2006) and of others suggests that students have some sound intuitions about data (Bakker, 2004). Teaching experiments by Cobb (1999, 2000) and Lehrer and Schauble (2004), have demonstrated how these intuitions can be refined and nudged towards reasoning that has inferential qualities. But there is little or no research that shows how teachers or students who begin as statistical novices come to a well-developed understanding of formal inference. In fact, many studies show that students have trouble mastering concepts, such as standard deviation, that are lynchpins of statistical inference.

In traditional approaches, statistical inference is introduced as a quantitative problem – usually of figuring out the probability of obtaining an observed result on the assumption that the null hypothesis is true. In this article, we lay out an alternative approach towards teaching statistical inference that we have been testing. This alternative involves delaying the problem of quantifying probabilities and instead laying the conceptual groundwork for inferential reasoning by focusing on a number of key properties of statistical processes (cf., Konold and Pollatsek, 2002). We have taken to referring to these key properties as "informal inference." We are aware of several other researchers who are exploring similar approaches (e.g., Arthur Bakker, Katie Makar, and Maxine Pfannkuch (2005), among others) and hope this paper will be part of a conversation that leads to more robust ideas of informal inference. We begin by describing our view of informal inference and then illustrate ways we have been trying to develop the component ideas in a recent workshop with teachers,

INFORMAL INFERENCE

At this point in our research, we describe "informal inference" as reasoning that involves the following related ideas:

I. *Properties of aggregates*. Aggregates of individual cases have emergent properties that are different from the properties of the individual cases themselves. It is these "aggregate" properties which we are interested in learning about from samples or batches of data.
   A. *Signal and Noise*. Two general types of properties of interest are signals (constant causes that are reflected in statistics such as the mean or line of fit) and noise (variable causes that serve to introduce variability around any signal).

B. *Types of variability*. In making judgments from a set of data about an underlying process, we must account for a number of forms of noise or variability.
   a. *Variability due to errors of measurement*. Part of the variability in data is due to errors of measurement that are present in any measurement we make.
   b. *Variability due to multiple causes*. Variability among outcomes of statistical processes arise from the interaction of a multitude of causal factors, many of which act independently of one another. The mean of a random sample from a process or population can be thought of as the net effect of all these factors. Natural variation within a population can also be attributed to the interaction of a multitude of causal factors.
   c. *Sample to sample variability*. Each time we randomly sample from a population or process, the sample looks different even if the population or process has not changed. Therefore, we cannot conclude that the process has changed simply because the current sample looks different than the previous one.

II. *Sample size*. Bigger samples are better (all else being equal) because they provide a more accurate estimate of population or process signals. As a sample grows larger, the mean of that sample tends to settle down close to the population parameter.

III. *Controlling for bias*. We randomly sample from a process or population to be sure not to introduce bias into the selection process and thus to increase the chance that the sample we get is representative of the population.

IV. *Tendency*. We can distinguish between claims that are always true and those that are often or sometimes true.

THE VISOR PROJECT

Over the past several years, we have worked with middle and high school teachers in the VISOR (Visualizing Statistical Relationships) project to learn about their evolving ideas of informal inference. The project consisted primarily of a series of seminars, attended by a group of teachers for 3 hours after school every other week. A central feature of our seminars was the use of two statistical visualization tools: *Fathom* and *TinkerPlots*; part of our research agenda was to study how the use of these tools affected teachers' developing understanding of statistical reasoning. The cohort described here included three middle school teachers and six high school teachers. Two of the high school teachers had taught A.P. statistics; the other four had taught a variety of math courses, but no or little statistics. This paper describes an episode from April 2004 in which these teachers explored a data set involving variability within a process. The basic story and data set were developed by Konold (2005). While the story the teachers read was a rich narrative, the basic facts are these:

> The Mus-Brush Company produces mushroom brushes, using a large machine whose output is on average 215 brushes every two minutes *if it is working normally*.
>
> If the electricity to the machine is interrupted, even for a brief time, it will slow down such that the output of the machine will be 10% lower on average.
>
> The Mus-Brush Company was robbed last night; in forcing the door open, the thief disrupted the electricity and the machine became less productive from that time on.
>
> There is a prime suspect who has an alibi between midnight and 3AM (he was seen at a bar), so the police have a special interest in determining if the break-in occurred before midnight or after 3, since the suspect has no alibi for those time intervals.
>
> We have data on Mus-Brush production every two minutes from 8PM until 5AM. Our job is to decide whether there is enough evidence to argue that the break-in occurred between 12 and 3, thus getting the suspect off the hook.

We chose these data and this analytic situation to focus in particular on the idea that by looking at sufficiently large samples, we can observe signals in fairly noisy data. We worked on this activity over two class periods during which the teachers conducted and discussed several different analyses. We focus here on one particular analysis, that of the series of means of Mus-Brush production ptm (per two minute) from 8PM until 5AM, taken over intervals of different sizes. The statistical arguments we present here are sometimes direct quotes from individual teachers and sometimes glosses of the general conversation involving several teachers. After describing the approaches teachers took using *TinkerPlots*, we offer some preliminary hypotheses about the conceptual issues that arose for the teachers.

TEACHERS' ANALYSES

Figure 1 is a *TinkerPlots* graph that many teachers created early in their analysis, in several steps. They first plotted the case number (the order in which the measurements were collected) on the X-axis and the output (the number of brushes manufactured in the past two minutes) on the Y-axis and then divided the points into hour-long bins. Each vertical bin is labeled by an hour, beginning with "eight" and continuing through "five." Each hour bin contains 30 points, where each point represents the number of brushes produced in a two-minute period during that hour. Within each bin, however, we have no information about the order in which the points were collected. Even before superimposing the colors that are part of Figure 1, most noticed that the output at the end of the night looked lower than that at the beginning of the night.

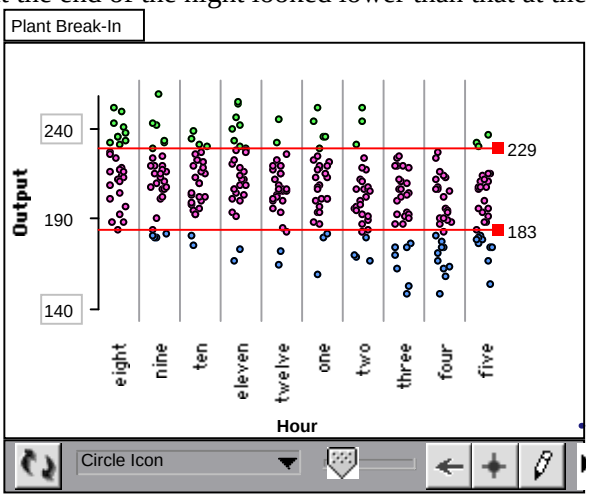

Figure 1: Machine output separated into hour bins, colored by low, medium and high output

Interestingly, the teachers appeared to have no difficulty understanding variability in this context, seeing it as natural variation in a mechanical process. This is in contrast to the case of a sample drawn from a population of individuals (e.g., 100 cats from the population of cats in a city) in which we have seen that the meaning of variability can be problematic. Teachers examining these data said things such as, "This machine is pretty variable!" and, as we will see below, used means extensively in their analyses. Indeed, this is one of the reasons we chose this particular context, as prior research has suggested that in data involving processes, students find it easier to think about both center and spread (Konold and Pollatsek, 2002; Lehrer, Schauble, Carpenter, and Penner, 2000)

In a second step that allowed them to think more precisely about the hour in which the break-in took place, teachers colored the points on the graph by "group," a derived variable included in the data set and based on the level of output of the machine for any given two minutes. Points that represent low output (less than 183) are colored purple, those with medium output are red (from 183 to 229), and those with high output (greater than 229) are green. (Lines on Figure 1 are included to emphasize the color borders, although they were not part of most teachers' graphs.)

One pair of teachers used a graph like Figure 1 to note that there were two hours in which there were no high outputs: three (from three to four) and four (from four and five) and that those two hours also had the most low outputs. This led them to hypothesize that something happened between two and three in the morning. They also noted, however, that the last bin – representing the time between five and six – had three high outputs and not as many low outputs as the two previous bins. This was initially confusing to them, because they knew that the machine, once operating at a lower level, would not resume normal operation until it was reset the next morning. However, because of the considerable variability in the machine's performance, they were willing to view the values in the last hour as possible even if the machine were operating at a lower level.

To investigate further their hypothesis that something happened during the hour from two to three (the hour labeled "two" in Figure 2), these teachers examined the means of each hour's production, looking for a "significant" drop between one hour and the next. In Figure 2, the mean Output for each hour appears as a blue triangle to the left of the corresponding bin. For example, the mean production for the eight o'clock hour is at the far left of the graph, at approximately 220. The points in this graph are colored by hour, rather than by lower, medium or high output as in Figure 1. Thus, all the purple points at the left of the graph represent production between eight and nine PM. On the graph in Figure 2, teachers focused on the means of hours one and two and noted that there was a large drop between them.

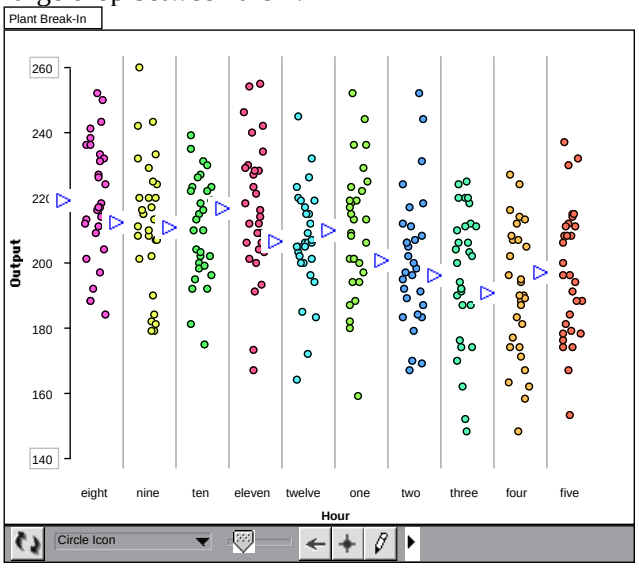

Figure 2: Means of each hour's production

Many of the teachers used a graph similar to that shown in Figure 2 to argue that the break-in occurred before 3, so the suspect was in the clear (since he was in the bar from 12 to 3). However, others disagreed, pointing out the drop from 11 o'clock to 12 o'clock (the red points to the aqua points). Could that be the break-in? Others noticed that the mean of the 2'oclock hour was 201 bptm, that of the 3 o'clock hour was 196 bptm and that of the 4 o'clock hour was 191 bptm. Perhaps the break-in happened between 3 and 4. Most of the teachers rejected the earlier time as indicative of the break-in because the production went back up significantly after twelve. To decide whether it was more likely that the break-in occurred between 2 o'clock and 3 o'clock or between 3 o'clock and 4 o'clock, teachers wanted to see a graph in bins of half an hour. Figure 3 is one graph they created.

We have added arrows in Figure 3 to indicate the means that several of the teachers came to perceive as indicative of a drop in production. The upper arrow points to the mean production between 2 and 2:30 (207 btpm), the lower arrow to the mean production between 2:30 and 3:00 (195 bptm). This change is larger than any other drop between two consecutive half hours, and the mean production continues to vary around 193 bptm, 10% less than normal production. The teachers therefore reasoned that the break-in most likely occurred around 2:30, at the boundary between the two half hours. The mean for the 2 o'clock hour shown in Figure 2 is 201 and that for

the 3 o'clock hour is 196. Using half hour bins, the teachers could see more detail in the graph. This graph suggests that the break-in occurred around 2:30, since the mean for 2:00 – 2:30 is so different from that of 2:30 – 3:00. (If the break-in occurred much earlier or much later than 2:30, the difference between the means of the two half-hours would not have been as large.) Comparing Figures 2 and 3, we can see that the more moderate change in mean from the entire 2'oclock hour to the entire 3 o'clock hour is due to the mixing of normal production and impaired production from 2 to 3. That is, while production drops from an average 207 bptm to an average 195 between the first and second halves of the 2 o'clock hour, looking at the hour as a whole produces a mean production of 201 mus-brushes per two minutes.

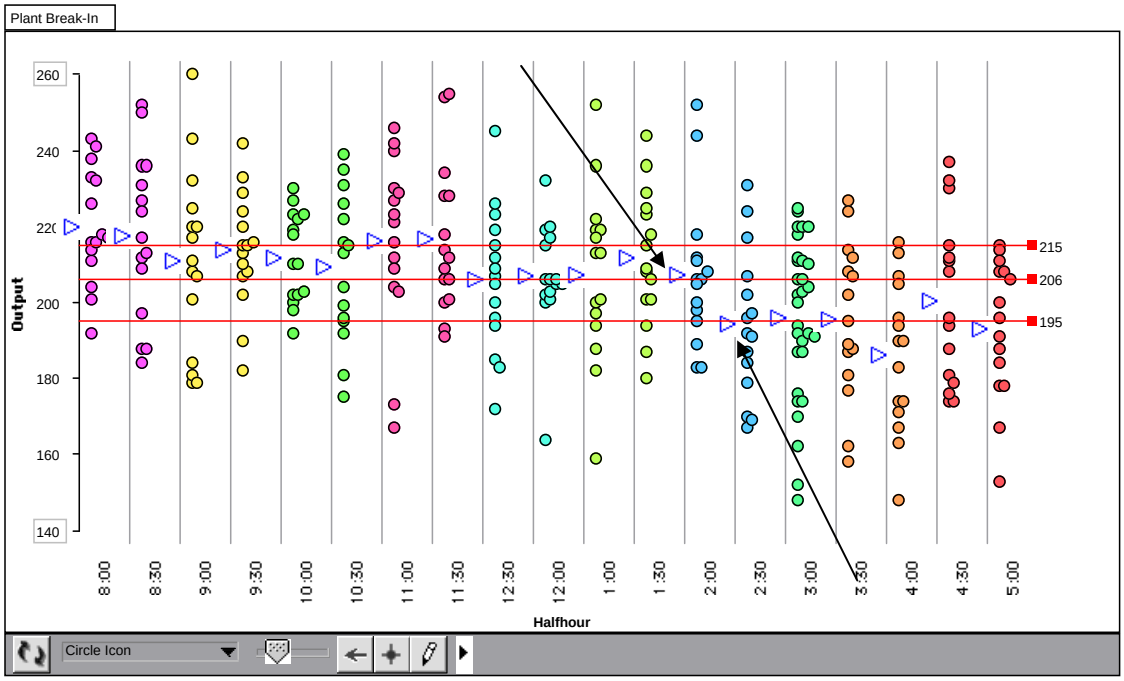

Figure 3: A drop in production between 2:00 and 2:30 is suggested by this graph

Alice (not her real name), the middle school teacher who created the graph in Figure 3, also added three horizontal reference lines, one at 215 (the machine's production when it is running normally), one at 195 (about 10% less than normal production), and one at 206 (the mean production for the entire night). She wrote this analysis of her graph:

Questions
1. at 11:30pm the plant's mean production begins to drop below the expected 215 brushes per two min. Does this signify a break in?
2. the actual mean for the evening is affected by the break in and is lowered to 206 bp2m, and seems steady during the 11:30pm to 2am time period.
3. then there is a second drop after 2 to 2:30am, where the fluctuation seems to move around a new mean of 195, which also is about 10% of the expected plant output. [Based on the rest of her analysis and on our conversations with her, we believe that she sees the break-in occurring between 2 and 2:30, not after that interval.]

Several of Alice's statements indicate to us that she has come to understand some key ideas that we view as part of a concept of informal inference. First, she appears comfortable with the idea that the *mean production for any half hour is likely to vary*; she notes that the mean "seems steady" between 11:30pm and 2am. In fact, the mean varies between about 206 and 217 during that period, but for Alice, that is within her expectations. We hypothesize that the fact that she was analyzing a *naturally varying process* made it possible for her to accept a certain amount of variability as "normal;" this is one hypothesis worthy of future research.

Second, Alice's analysis, unlike the previous ones we described, *includes the absolute position of the means as well as their relative position*. That is, rather than just looking at a "drop" between two consecutive means, she notes when the mean value for a half-hour begins to "move around a new mean of 195," which she identifies as the plant output after the break-in.

Third, she notes correctly that the mean production for the entire evening is about 206 bptm, but incorrectly sees the relative steadiness of the mean around 206 between 11:30 and 2:00 as consistent with this fact. Instead, the mean of 206 is derived from a mean of 215 before the break-in and a mean of 195 after the break-in. In fact there is no time at which the machine is producing 206 bptm. Alice's error might be related to the difficulty people have in general *accepting the mean of a sample as a value that is quite likely not the value of one of the individuals in the sample.*

As for the role of the *TinkerPlots* software in this analysis, it is hard to imagine this many different graphs being generated and compared by hand. In particular, the abilities to look at data in bins of different sizes (Rubin and Hammerman, 2006; Hammerman and Rubin, 2004) and to use both visual and numerical judgments (Rubin, Hammerman, Campbell and Puttick, 2005) appear to be necessary parts of the analysis. Future research should focus on describing the role of more specific aspects of *TinkerPlots* in this kind of inferential analysis.

In sum, we believe that these analyses suggest important topics to explore in more depth as we build a description of informal inference and a theory of how people learn its underlying principles.

ACKNOWLEDGMENT

Support for this paper comes from the National Science Foundation (NSF) under grant REC-0106654, Visualizing Statistical Relationships at TERC, Cambridge, Massachusetts, USA. The views expressed are those of the authors and do not necessarily reflect those of the NSF.

REFERENCES

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. Doctoral dissertation, Utrecht University.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning,* 1(1), 5-43.

Cobb, P. (2000). Conducting classroom teaching experiments in collaboration with teachers. In A. E. Kelly and R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education,* (pp. 307-333). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P., McClain, K., and Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction,* 22(1), 1-78.

Hammerman, J. K. and Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal,* 3(2), 17-41.

Konold, C. (2005). *Exploring data with TinkerPlots*. Emeryville, CA: Key Curriculum Press.

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

Lehrer, R., Schauble, L., Carpenter, S., and Penner, D. (2000). The inter-related development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel, and K. McClain (Eds.), *Symbolizing and Communicating in Mathematics Classrooms: Perspectives on Discourse, Tools and Instructional design,* (pp. 325-360). Mahwah, NJ: Lawrence Erlbaum Associates.

Pfannkuch, M. (2005). Informal inferential reasoning: A case study. In K Makar (Ed.), *Proceedings of the International Forum for Research in Statistical Reasoning, Thinking and Literacy*. Auckland, NZ. Brisbane: University of Queensland.

Rubin, A. amd Hammerman, J. K. (in press). Understanding data through new software representations. In G. Burrill (Ed.), *Thinking and Reasoning with Data and Chance—2006 NCTM Yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Rubin, A., Hammerman, K., Campbell, C., and Puttick, G. (2005). The effect of distributional shape on group comparison strategies. In K Makar (Ed.), *Proceedings of the International Forum for Research in Statistical Reasoning, Thinking and Literacy*. Auckland, NZ. Brisbane: University of Queensland.