INTERACTIVE VISUALIZATIONS OF STATISTICAL RELATIONSHIPS: WHAT DO WE GAIN?

Andee Rubin
TERC
USA

This paper describes a project that involves statistical researchers and software designers in a collaboration designed to accelerate progress in research on statistical thinking and the development of effective software tools for statistical education; the two tools in question are TinkerPlots and Fathom. Most of the paper is devoted to a description of teachers analyzing a dataset first without, then with technology and to a discussion of the implications of such observations for both research and software development.

INTRODUCTION

Data literacy has become a fundamental skill for living in a democratic society. Data-driven decision making is increasingly at the heart of health care, education, public policy, business, and environmental activities. Many scientific statements construed as "fact" are based on correlations derived through statistical analysis of large amounts of data. As statistics educators, we know that most people are woefully bad at understanding data beyond the simple (and often misleading) pie graphs and bar graphs they see in the newspaper. However, there is currently some reason for optimism. Two recent developments in statistics education provide the opportunity for significant advances in helping non-statisticians judge statistical claims.

1) Research in statistical thinking has begun to yield models of people's conceptions that are detailed enough to have practical, pedagogical implications;
2) Powerful new software tools designed explicitly for statistical education provide new visualizations with enormous potential for making statistical thinking accessible.

This paper describes a collaboration between researchers and software developers that will accelerate progress *both* in the research literature on statistical thinking and in the development of effective software tools for statistical education. The collaborators are Andee Rubin at TERC in Cambridge, MA, a statistics education researcher; Bill Finzer at KCP Technologies, a software designer and lead developer of Fathom a statistical education tool used primarily in high school (Finzer, 2001) and Cliff Konold at the University of Massachusetts, statistics education researcher and designer of Tinkerplots, a statistics construction tool for middle school students (Konold & Miller, 2001).

RESEARCH QUESTIONS

Our research focuses on the effects of interactive visualizations on teachers' and students' reasoning about statistical association; this paper will address the following two questions:

Research Question 1—Examining shapes of univariate distributions: What features of univariate distributions are salient for teachers; how do they use these to describe such distributions? How does technology affect teachers' perceptions of data; can it shift their attention from individual data points, or cases of similar value, to seeing an entire group and the characteristics of that group?

Research Question 2—Comparing univariate distributions: What approaches do teachers take to group comparisons? Do they use group characteristics (e.g. the mean), argue based on "cut points," or focus on outliers? What aspects of a distribution influence these choices? How does software support and influence these comparisons?

This paper looks at these two questions in the context of a data study group of 11 teachers who meet every other week for three hours throughout the school year. Teachers in the group are from both middle and high schools in both urban and suburban school districts. They vary widely in their experience with data and computers. In each session, the group analyzes a data set—sometimes collected in the session, sometimes provided—on paper and/or using either Tinkerplots or Fathom.

THE BATTERY DATA

The analysis discussed in this paper is from the third session of the teacher group. In the first two sessions, teachers collected data about themselves and analyzed it both on paper and using TinkerPlots. In this third session, we gave them a simple dataset originally used by Cobb (1999) that comprised lifespans in hours for two fictitious brands of batteries: Tough Cell and Always Ready. There were pieces of data for each battery brand, as presented in Table 1.

Table 1
Lifespan in Hours for Two Fictitious Brands of Batteries

BATTERY	LIFE		BATTERY	LIFE
Always ready	97		Always Ready	43
Always Ready	115		Tough Cell	84
Tough Cell	98		Tough Cell	104
Always Ready	73		Always Ready	111
Tough Cell	109		Always Ready	108
Tough Cell	103		Always Ready	113
Tough Cell	106		Tough Cell	115
Always Ready	115		Tough Cell	91
Tough Cell	121		Tough Cell	101
Always Ready	115		Always Ready	115

The question posed to the group was:
If you had to choose one of these brands of batteries, which one would you choose? Come up with a way to represent the data that supports one answer to the question.

In pairs or groups of three, teachers spent about 20 minutes working with the data on paper, then moved to the computer to analyze the data using TinkerPlots. At the end of the session, they presented their conclusions to the rest of the group.

An important aspect of the problem is that the distribution of lifetimes for each kind of battery has different shapes. The Always Ready batteries have a skewed distribution with a significant mode at 115, but two low values at 43 and 73. The Tough Cell batteries have the highest value (121), but a more symmetrical distribution with a clump between 98 and 109; the lowest value was 84. In this paper, I describe the way one set of three teachers approached the data, both as they began to analyze the data on paper and then as they moved to TinkerPlots. The focus of the description is on the roles of central tendency, variability, group characteristics, and individual points in the teachers' analysis.

ANALYZING THE DATA USING PAPER AND PEN

The first thing the three teachers, Grace, Karen and Candace (these are pseudonyms) did was to calculate a mean lifetime for each battery. By this measure, Tough Cell was "better"; its mean was 103.2, compared to the Always Ready, which was 100.5. While this is a group characteristic, and, in fact, the one we are generally told to use for comparing two distributions, it seemed that these teachers calculated the mean as an automatic first step, but did not linger on whether the difference between means was "significant". Instead, they immediately began to talk about the 43 value, which they identified as an outlier. The presence of the outlier seemed to mean to them that the comparison of means was not the best way to make a decision between the brands—or was, at best, incomplete.

The next part of the conversation consisted of a detailed discussion of the appropriate approach to the outliers in these data. There seemed to be several influences on their decision-making:

1) At this point, the perception of outliers was as points to be eliminated—exceptional "duds" that had to be disposed of to get to the "real" distribution—rather than representing 10% of the batteries that should be noted.

2) Some of the teachers had a "fairness" principle: if you delete an outlier from the bottom of the distribution, you should delete one from the top "to be fair".
3) The repeated 115s in the Always Ready distribution were particularly salient.

Sometimes these influences conflicted. For example, Karen suggested getting rid of the 43, but wanted to get rid of one at the top of the Always Ready distribution. However, the high values of this distribution were all 115 and Karen was reluctant to delete one of them. She admitted that she was being influenced in her decision about deleting another point by the rest of the distribution—even as she thought she "should" have been making these decisions "blindly," without considering the actual value of the matching outlier. One interpretation of this reluctance is that the 115s must have been "right" because there were so many of them, but the 43 might have been a "wrong" value. Karen recognized the difficulty of making these outlier decisions without a picture of the data and thought a box plot, where she had a visual sense of outliers, would have been helpful.

The teachers had a similar dilemma with the 84 in the Always Ready distribution. It too was far away from the modal clump of the data, so was a candidate as an outlier. But deleting this data point suggested that the highest data point should be deleted as well and, as this was the highest overall value, deleting it seemed to get rid of an important piece of information. During this part of the discussion, the idea of comparing means was secondary to that of identifying outliers. A few times, the three teachers considered what the effect of deleting these points would be on the mean by noting how far away they were from the current mean, but this was not the focus of their discussion.

ANALZYING THE DATA USING TINKERPLOTS

The teachers next proceeded to look at the data using Tinkerplots. We were interested in what effect the visualizations available in Tinkerplots would have on their view of the distributions. The main move they made with the software was to "separate" the variable "life" into six bins (kind of a histogram, but without the "stacking" of points). This separation resulted in Figure 1, in which Always Ready batteries are represented by solid circles and Tough Cells by the dotted ones. Notice that in this graph all of the points except for four are in the two bins on the right. Of the four that are outside of this clump, two are Always Ready and two are Tough Cells. The teachers noted this strong symmetry – two data points of each type outside of a modal clump – and used it to argue that it was appropriate to consider the two lowest values for each battery type as outliers and, thus, to delete them. Their comment on the graph: "looking at this sort, I would throw out lowest two of each type" reflects this interpretation. Notice at this point that the necessity to delete equal numbers of points from the top and bottom of a distribution is no longer observed. Instead, the symmetry of deleting the same number of points from each distribution seems more convincing.

DISCUSSION OF THIS EPISODE

The preliminary analysis of this episode illustrates how our work can affect research in both statistical thinking and software design. As statistics educators, we note that, while these teachers calculated the mean for each distribution, they were dissatisfied with this measure because of the shape of the distributions – in particular, several distinct outliers. This leads us to ask (and, hopefully, to answer) additional questions: How does this tension between central tendency and spread (in the form of outliers, here) play out in different distributions? How is it affected by the number of points in a distribution? What affects teachers' interpretation of outliers as either points to be disregarded – or points to consider carefully (e.g. would you want to buy a battery if 10% were significantly worse than any others?)

In terms of software design, we note that the representation the teachers constructed in Tinkerplots made it "clear" that there were four outliers—yet a different set of bins might have prompted a different conclusion. One hypothesis that bears consideration - and further research - is that this hybrid representation, with characteristics of both histograms (binning) and dot plots (individual points represented), may have interesting and unexpected influences on users' conceptions of distributions, in particular, their identification of outliers. Observing this connection prompts us to ask if the software should behave any differently, if so, how and if not, what implications there might be for instructional materials to accompany the software.

ISSUES FOR THE FUTURE

In addition to these specific observations, this episode suggests three major themes that we anticipate will thread their way through much of our work:

1) The role of context. In their discussion, the teachers almost always referred to the data as batteries; that is, their thinking was strongly rooted in the context of the data. We conjecture that their statistical reasoning may have changed had the context been different. Being able to talk about the interplay between distributional shape and context productively will be one challenge of our work.

2) The role of personal experience. One of the striking parts of the analysis above is the teachers' notion of fairness: deleting one outlier from the bottom of a distribution necessitated deleting one from the top. It may be important that one of the calculations teachers often do is to "average" a student's grades and may throw out both the lowest and the highest score before taking the average. While we don't know that there is a connection between the way these teachers compute grades and the way they analyzed the batteries data set, there will no doubt be cases where people approach data differently because of their experience and background

3) Even in our early research, it has become clear that the notion of "sample" is unavoidable. Teachers' analysis of the batteries data differed depending on how they viewed the twenty data points in terms of a sample. Some groups seemed to think that these data were representative, e.g. that one very low number out of 10 would suggest that 10% of the batteries in the larger population might be duds. Other groups were more willing to disregard a low value, as if it could be regarded as an anomalous value in terms of the larger population as well.

These three themes are bound to enrich and complicate our work, in ways that will bring us closer to a better understanding of statistical reasoning and its implications for software design.

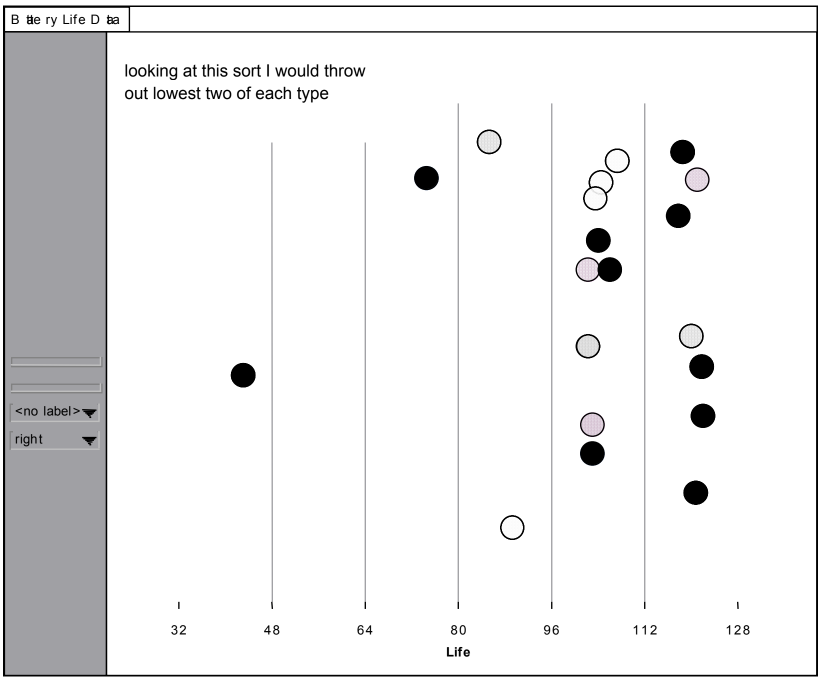

Figure 1. A look at the Data Using Tinkerplots.

REFERENCES

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

Finzer, W. (2001). *Fathom Dynamic Statistics* Software.* Key Curriculum Press.

Konold, C., & Miller, C. (2001). *Tinkerplots version 0.42. Data analysis software for the middle school*. Amherst: University of Massachusetts.