

WHAT DOES DRAGGING THIS DO? THE ROLE OF DYNAMICALLY CHANGING DATA AND PARAMETERS IN BUILDING A FOUNDATION FOR STATISTICAL UNDERSTANDING

William Finzer
KCP Technologies, United States
bfinzer@keypress.com

Dynamic manipulation of mathematical objects in a computer-learning environment allows a learner to build an intimate, visceral relationship with those objects. Dynamic manipulation is direct and continuous. Dragging data allows the learner to experience the affect of data change on statistical measures and their visual representations such as the vertical line and the computed value in the plot above. A systematization of the very large set of opportunities for dragging data and the kinds of learning fostered is presented.

INTRODUCTION

Imagine that children were taught to ride bicycles that had, instead of handlebars, a calculator-like keypad into which they typed the angle in degrees to which they wished the wheel to turn. (And remember to press Enter!) Would anyone learn to ride?

Immediate, intimate, visceral feedback is accepted as essential to much learning. Why do we implicitly assume that it is *not* essential to learning data analysis and statistics? If all bicycle steering were accomplished through keypads, a few “expert” cyclists would emerge to amaze the rest of us cycling “illiterates.” It may be that as controls for the exploration of data continue to evolve, we will find that the barriers to entry and expertise become much lower. Such is the hope underlying this paper.



For more than twenty years, using a mouse to drag things on computer screens has become an increasingly popular pastime. We are talking here about a subset of dragging things we choose to glorify with the label “dynamic manipulation of mathematical objects” (Finzer, 1998). This manipulation is *direct* in that the thing you wish to change is the thing you drag. It is *continuous* in that changes take place during the drag, and everything else on the screen that depends on the object being dragged also changes. Finally, the environment in which you are doing this is *immersive*, that is to say, it draws you in so that you feel part of it, with the computer interface only minimally getting between you and the achievement of your mathematical goals.

This paper describes the potential for dynamic manipulation to foster increased understanding of statistical concepts. The author draws on his experience as a software developer and as a teacher; i.e., the claims made are speculative rather than based on research. Also, though the examples are based on usage of *Fathom* (Finzer, 2005), there are other data analysis environments—for example *TinkerPlots* (Konold, 2005)—that have similar dynamic manipulation features. The general plan of the paper is to present examples of dragging and, for each, to draw attention to the kinds of concept building that could be taking place for the learner so engaged.

THE MEAN

We begin in the middle; that is, with the mean. A rich literature surrounds the learner’s conceptualization of the mean. (Mokros, 2000; McClain, 2000).

What happens to the mean when data are changed? For example, in Figure 1, each dot in the plot represents one mean salary of full professors at a single university or college in the U.S. As one point is dragged, the mean changes both as a numerical value shown below the plot and as a vertical line drawn in the plot.

Consider various actions the learner can execute. For each action we list possible realizations the learner could have.

- *Dragging a single point:* The mean always changes, no matter which point. The change is always in the same direction as the direction the point is dragged. The change in the mean is always less than the change in the point (except when there is only one point). The change in the

mean is always $1/N$ of the change in the point, no matter which point it is; therefore an outlier has the same “influence” as any other point.

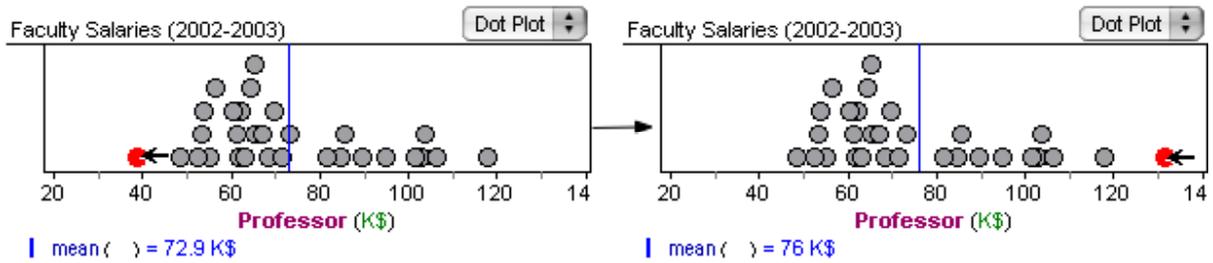


Figure 1: The lowest value in a dot plot is dragged to become the highest value

- *Dragging points sequentially:* Moving one point a certain distance in one direction can be balanced by moving any other point the same distance in the opposite direction.
- *Dragging multiple points:* The influence of multiple points is proportional to the number of points being dragged. The influence of *all* the points together is one. Dragging $N-1$ points to the right has the same relative effect as dragging 1 point $N-1$ times as far to the left.
- *Adjusting multiple points and considering the position of the mean relative to the distribution:* For a symmetric distribution, the mean is necessarily in the middle. The mean can never lie outside all the values. The mean may be made to be arbitrarily far from any of the actual data values. If there is a single model clump and there are no extreme outliers, the mean will lie within the clump.

All measures are susceptible to this kind of characterization. It is instructive to compare the mean with the median.

<i>Mean</i>	<i>Median</i>
Always changes, no matter which data point you drag.	Only changes when the middle one or two points are dragged.
The change in the mean is always $1/N$ of the change in the point, no matter which point it is; therefore an outlier has the same “influence” as any other point.	The amount of change in the median is either zero, half the change in the point, or equal to the change in the point.
A single point can, by itself, determine the value of the mean.	Only one or two middle points determine the value of the median.

WHAT CAN I DRAG AND WHAT CAN I LEARN?

Translating an Axis Scale

Let’s start with something apparently straightforward—changing the scale of a graph axis through translation with no data present. In Fathom the user can do this by clicking anywhere in the middle third of an axis and dragging in either direction. Could anything be simpler?

We postulate that through repeated encounters with axis translation a student absorbs certain useful ideas. A student’s inner dialog might sound like this: “The axis number line goes on forever in both directions. Any one part is just like another.” These would not be *new* ideas to a student, but encountering them again while working with data can help students more deeply understand a powerful data analytic concept: The position of data values on a number line is largely irrelevant without an external context, a measuring rod of some sort that gives meaning to the numbers. The numbers 198, 213, and 216 by themselves have no meaning. Knowing that they are heights of people measured in centimeters brings them into focus.

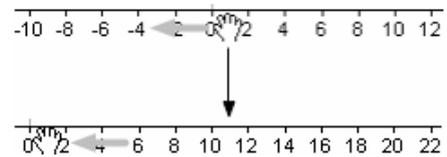


Figure 2: Dragging the middle portion of an axis scale translates it

The act of dragging an axis scale is so related to transforming data that it may help build the foundation on which an understanding of the process of transforming data comes to rest. But another act of dragging gets more directly to the heart of transformation.

Translating Data

Normally, applying a translation to data would be done by formulaically computing the transformed values in a new attribute. To center the transformed data on zero, for example, we might use the formula $Height - mean(Height)$. But in a dynamic manipulation environment, the learner can select all the data points and drag them to a new position approximately centered on zero. While in the long run this may not be good data analysis practice to encourage, as an introduction to data transformation it gives the student a chance to notice that the shape of the distribution remains constant and to explain this constancy as a consequence of the fact that the data values all move together, maintaining their relative positions.

Understanding that a linear transformation is reversible comes naturally from the realization that the data can be dragged back to where they started. Watching the data move as we drag produces a certain knowledge that nearly all of what is important about the data values remains intact and that if the amount of the drag is known, the original data can be reconstituted.

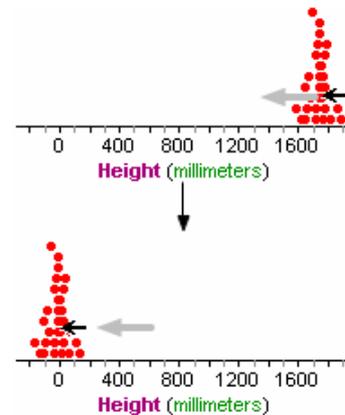


Figure 3: Dragging all the data values accomplishes a translation by a constant

The Shape of a Histogram

A central task confronting a learner in data analysis is that of becoming comfortable with talking about and reasoning with distributions. A key component of distributional thinking (Rubin, 2005) is dealing with the *shape* of a distribution. Histograms are frequently the graph of choice for displaying a distribution, but they have a quirk that can catch learners unaware. Consider the uniform distribution shown in the dot plot of Figure 4. The middle graph, a histogram, does as well as the dot plot does at revealing the uniformity of the distribution. But as the user drags a bin boundary, increasing the width of the bars, the histogram dramatically and surprisingly changes shape, and no longer looks uniform at all!

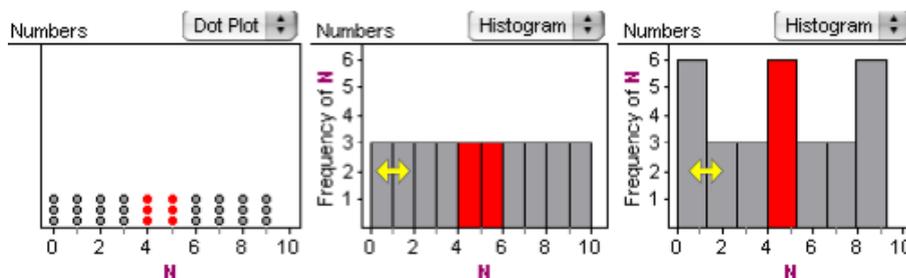


Figure 4: Three views of a uniform distribution. The selection of cases with values 3 and 4 shows the mechanism by which the histogram becomes nonuniform as the bins get wider

Does this “aliasing” quirk of histograms hinder learners in their efforts to establish a stable concept of a distribution’s shape? Does the ability to dynamically change a histogram’s bin width help learners come to see that there is an underlying shape that persists across a range of bin widths? Consider an example. The distribution of 500 people drawn from the year 2000 U.S. Census illustrates a typical distribution question: Does the bump in the distribution around age 45 represent baby boomers, or is it an aliasing artifact? As learners drag the edge of a histogram bin and experience the persistence of the boomer cluster, we hope they are learning to focus on shape and discount spurious artifacts.

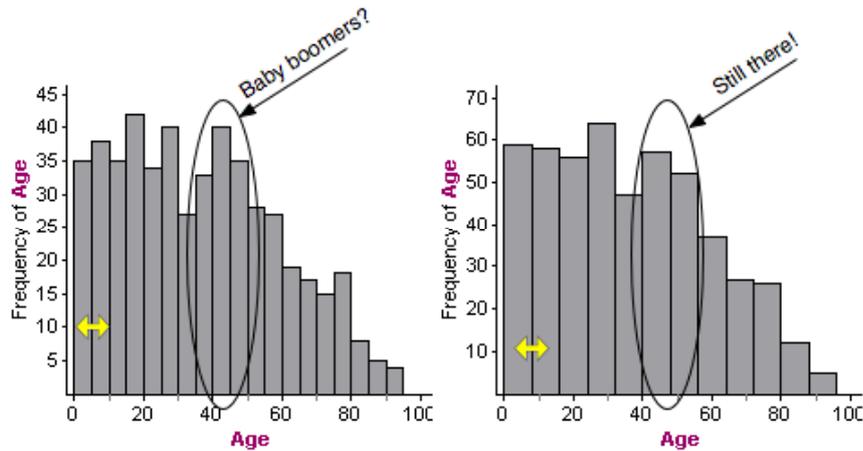


Figure 5: Continuously changing the histogram’s bin width shows that the baby boomer “feature” is stable

Movable Lines

It is pedagogically useful to provide users with tools for “eyeball” estimation of distribution statistics, and movable lines (either univariate or bivariate) are important examples of such tools. Figure 6 shows the one-dimensional case in which the user is estimating the “center” of a distribution. In Figure 7, the user has placed two movable lines in a bivariate distribution to estimate the slopes of each of two subgroups of points.

As the user drags a movable line, it moves across the distribution and the one or two values that define it update. There are many ways learners can use movable lines. In reading the statements that follow, imagine the user dragging the movable line, the word “*here*” indicating a particular placement of the line.

- About *here* is the axis of symmetry.
- *Here* is the expected mean, but the actual mean is much lower.
- There seem to be two groups, one *here*, and the other *here*.
- I’ll put the movable line on top of the least squares line and the drag an outlier. I can easily see how much influence the outlier has by the change in the least squares line compared to the movable line.
- The edge of the points in the scatter plot is about *here*.

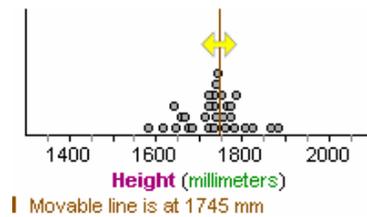


Figure 6: A distribution of heights with a movable line

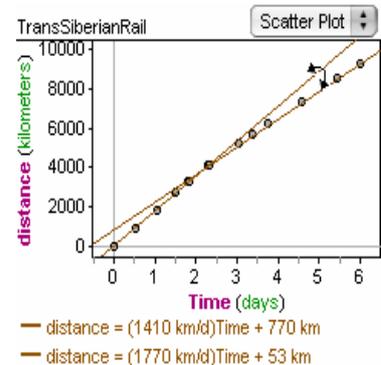


Figure 7: Two movable lines help describe a train trip across Russia

Sliders as Draggable Parameters

A slider is a named value whose magnitude can be changed by dragging. Sliders most often function as parameters in a model. In Figure 8, sliders determine the mean and standard deviation of a population from which a sample is taken. Dragging a slider causes samples to be drawn from the newly defined population.

The movement of the slider itself holds little interest, but the *effect* of that movement, allows the learner to observe a continuous change in a model. In the standard deviation example, the learner experiences a great many samples in rapid succession, each with slightly different spread. This experience of repeated sampling must, we postulate, help build a foundation that leads to a better understanding of data analysis and inference.

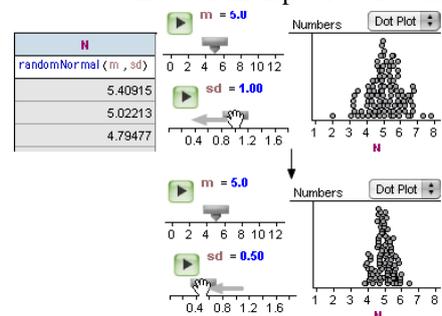


Figure 8: Sliders *m* and *sd* control the mean and standard deviation of 100 numbers from a normal distribution

Curve Fitting with Sliders

Sliders serve extremely well as coefficients in plotted functions, allowing the learner to gradually adjust the shape of a curve and experience the commonality of functions belonging to a family. Automatic curve fitting may serve experienced practitioners well, but “eyeball” fitting by dragging gives beginners a chance to interact with the model they are building. And, as in the example of Figure 9, it allows fitting of features such as the edges of point clusters.

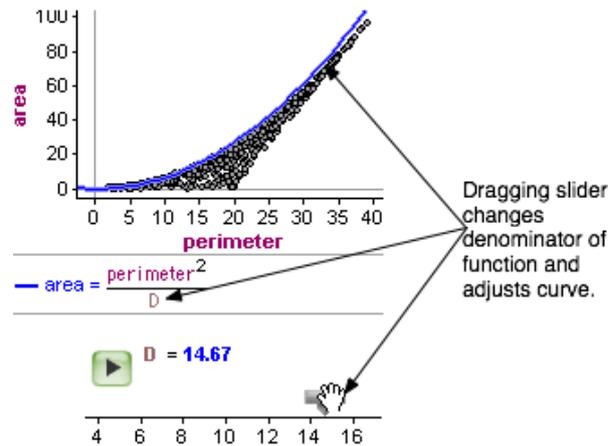


Figure 9: Dragging a slider allows the user to adjust a curve to fit a feature of a scatterplot

Sliders as Input to Inference Objects

A slider can serve as a parameter to an inference; e.g., a *t*-test, as shown in Figure 10. Dragging the slider causes computed quantities in the inference and any derived displays to update. Eavesdropping on the learner’s inner dialog:

- The standard deviation doesn’t depend on count. Is that because it’s an estimate of something in the population?
- The *p*-value and the shaded area under the curve vary together. Bigger *n*, smaller *p*-value.
- For small values of *n*, the shape of the *t*-distribution changes, too. But once $n > \sim 10$, the curve doesn’t seem to change at all. For small *n*, the curve is lower and more spread out.

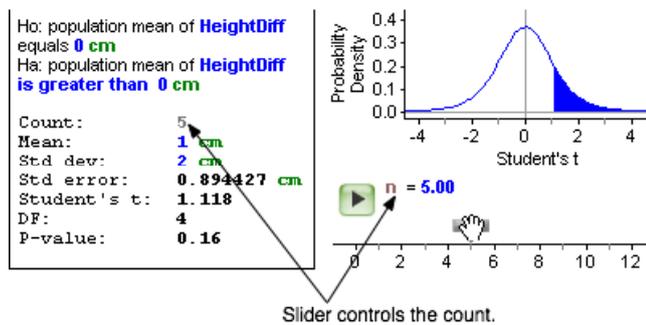


Figure 10: A slider allows the user to dynamically explore the effect of sample size on a *t*-test

Dragging Data and the Effect on an Inference

We began by looking at the effect of dragging data on one of the simplest of measures—the mean. We end by looking at its effect on statistical tests of significance. For the data shown in Figure 11, we have two groups of values, G1 and G2. We can compare the mean of one group with the mean of another, or, if we can treat the values as paired, we can test whether the mean of the differences is significantly different from zero. Both tests are shown. Notice the *p*-values.

Now, we reach in and drag the values of G1 until the greatest and least values are approximately interchange, resulting as shown in Figure 12.

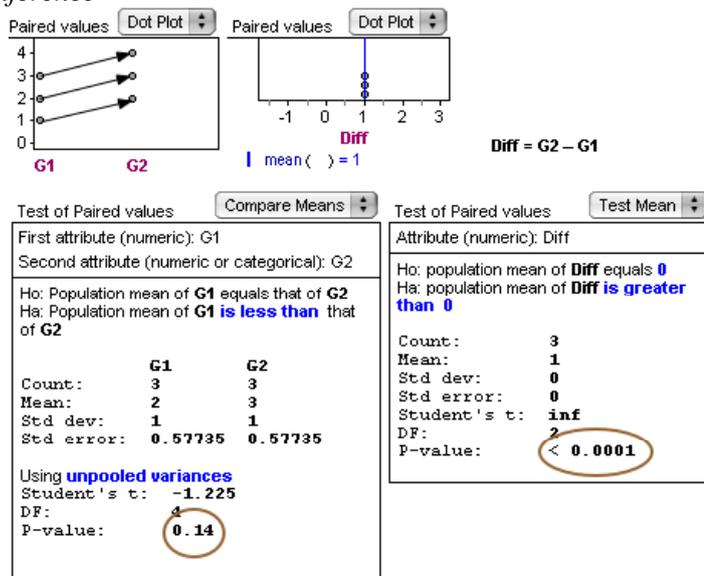


Figure 11: In a fictitious dataset with three cases, each of which represents a pair of measurements, the difference of means is not significant, but the mean of the differences is significantly different than zero

It's informative to watch the p -values change during this drag. At some point the p -value for the significance of the mean of the differences becomes greater than that for the difference of means. Most of us, accustomed as we are to believe that a paired t -test is more sensitive than an unpaired t -test, feel somewhat perplexed by this result. Going back to dragging the data gives us a manipulative tool to help solve the puzzle. For the author, the necessary insight came from watching the spread of the values of the differences increase.

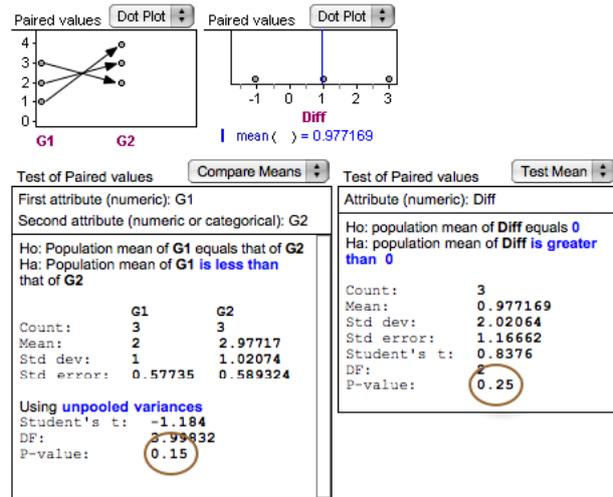


Figure 12: With G2's max and min values interchanged, the p -value of the paired t -test changes dramatically

BUILDING A FOUNDATION FOR UNDERSTANDING

One important difference between learning to ride a bicycle and becoming statistically literate is that while the kinesthetic ability suffices for the former, statistical literacy rests on cognition. Few cyclists can explain *how* they stay upright, but explanation lies at the heart of data analysis. What about the kinesthetic experience of dynamic manipulation helps build a foundation for statistical understanding?

- *First is the multiplicity of frames.* Recall the situation of Figure 8 in which dragging a slider caused new samples of numbers drawn from a normal distribution whose standard deviation was determined by the slider value. In fifteen seconds of manipulation, the learner sees perhaps fifty sample dot plots. We imagine a visual storehouse from which the learner can draw in the future, not exact images, but impressions of shape and, most importantly, variability. When confronted with a particular set of sample values, the learner can compare it with past visual experience and decide whether the distribution from which it came was likely normal, or whether an outlier deserves more than passing notice.
- *Second is the association of magnitude and direction of the drag with the magnitude and direction of the effect.* Consider the observation, “When a point in a dot plot is dragged to the right, the mean moves to the right, but less far than the point was moved.” With repeated experience dragging points, we suppose that the association of change in mean with change in data value becomes *intuitive*. Most of what we mean by a “foundation” for understanding is the accumulation of intuitions, the set of things we know without having to think deeply about them.
- *Third is the effect of immersion.* Dynamic manipulation draws people in—to explore, to play, and to ask “what if.” The more time you spend working with data and statistical objects, the more you will become adept and the stronger your foundation will become for continued learning.

REFERENCES

Finzer, W. and Jackiw, N. (1998). Dynamic manipulation of mathematical objects. White paper presented to the NCTM 2000 Electronic Format Group, http://www.keypress.com/sketchpad/general_resources/recent_talks/nctm_standards2000/index.php.

Finzer, W. (2005). *Fathom Dynamic Data Software*, Key Curriculum Press.

Konold, C. (2005). *TinkerPlots Dynamic Data Exploration Software*, Key Curriculum Press.

McClain, K., Cobb, P., and Gravemeijer, K. (2000). Supporting Students' Ways of Reasoning about Data. In M. Burke (Ed.), *Learning Mathematics for a New Century, 2000 NCTM Yearbook*, (pp. 174-187). Reston, VA: National Council of Teachers of Mathematics.

Mokros, J. and Russell, S. J. (2000). Children's concepts of average and representativeness. Working paper published by TERC, Cambridge, MA.

Rubin, A., Hammerman, J., Campbell, C., and Puttick, G. (2005). The effect of distributional shape on group comparison strategies. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4)*.