DATASPACE—A COMPUTER LEARNING ENVIRONMENT FOR DATA ANALYSIS AND STATISTICS BASED ON DYNAMIC DRAGGING, VISUALIZATION, SIMULATION, AND NETWORKED COLLABORATION

William F. Finzer and Timothy E. Erickson, Key Curriculum Press, Berkeley, USA

Four critical ingredients in the design of software for teaching and learning statistics are considered in the context of the DataSpace project at Key Curriculum Press. Dynamic dragging lets students see the transition from one state to another. Statistical visualization gives students ways to understand difficult concepts. Simulation, integrated into the learning environment, models hypothetical statements prevalent in statistics, and gives students a tool for doing inference through resampling. Networked collaboration makes it possible for students to gather data relatively painlessly, to share data and ideas with others, and to gain access to the wealth of timely data available on the Internet.

Technology should give learners tools so they can construct their own conceptual understanding. In the domain of statistics, we believe this means students should be able to experiment with ways of visualizing and analyzing data so that they can ask and answer questions of the form, “What will happen if …?” What will happen to this display if there are outliers? What will happen if we try to use our model with the most recent data? What will happen to our measure of difference if we repeatedly reshuffle and resample the given data? With support from the National Science Foundation[1], the DataSpace project at Key Curriculum Press has been developing a computer learning environment in which secondary and lower division college students may experiment in these ways.

Here we discuss four aspects of DataSpace’s design that contribute to its usefulness in constructing conceptual understanding: dynamic dragging, visualization, simulation, and networked collaboration.

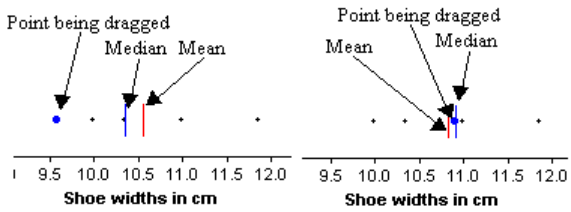
Dynamic Dragging

In our previous work with *The Geometer’s Sketchpad*[2], we observed that a great deal of the insight students gain about geometry comes from being able to see the changes that take place in a geometric construction while dragging some given of the construction. There is a qualitative difference between comparing two disparate states of a system and actively controlling the transition between those two states. We summarize this observation by saying, “The learning takes place during the drag.”

The difference between a static geometry and one that can be actively manipulated, a *dynamic geometry*, is profound. In our research in the DataSpace project, we have been

surprised and delighted to discover that active manipulation of statistical objects on the computer screen, a *dynamic statistics*, has a similar, profound effect on ease of learning.

Here we are cursed by the printed page, and we will have to ask you, the readers, not only to imagine what it is like to see what we describe as graphs *in motion*, but also to imagine what it is like to *control* the dynamic display. Among the statistical objects that can be changed through dragging on the computer screen are data values in plots, values of model parameters, scales on axes, bin sizes, and positions of lines.



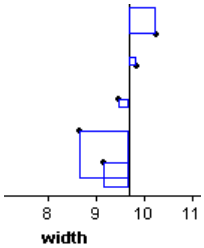
Consider learning about mean and median. Each point in the plots at right shows the width of one of five shoes measured at its widest point. The vertical lines show the positions of the mean and median for the five data values. A student selects and drags one of the data points, observing the behavior of the two lines. The mean line moves continuously. The median line does not move at all except while the dragged point is the middle point of the five.

Students have to explain the observed difference in the behaviors of the two lines in order to make sense of what they are seeing. Under what conditions is the mean to the right of the median, to the left of the median, or on the median? What happens if we add another data point, drag two data points at once, or drag all the data points at once? If we drag one of the points one unit, how far does the mean line move?

Now let us give students tools with which they construct their own measures of center. They define their measure algebraically so that it gets plotted along with the mean and median and responds dynamically to changes in the data. By comparing their own measures with traditional measures students can come to understand that statistics are inventions of the human mind, not magic and mysterious incantations.

Visualization

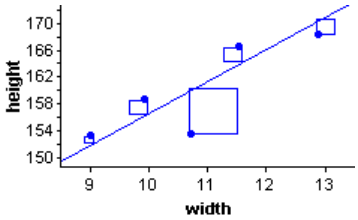
Another way to provide students with tools to understand statistical concepts is give them *visualizations*. These are dynamic diagrams that show how a concept works, and help reveal the meaning of the underlying (and often opaque) mathematical symbols.



Suppose we spread vertically the five shoe width points and place a moveable vertical line among them as shown at right. From each point we construct the residual between that point and the line as a segment from the point to the line and perpendicular to the line. Finally, we use each residual as the side of a constructed square.

The moveable line is a model of our data and the sum of the areas of the constructed squares is one measure of the degree to which the data conform to the model—the smaller the measure, the better the model. For what position of the line do we get the least sum of squares? As students drag the line, they see the squares change size and a computed sum go through a minimum. When they place the mean line in the same plot, they see that when the sum of squares is a minimum, the moveable line lies exactly on top of the mean. So we have a dynamic visualization of the rather deep concept that *the mean is a least squares model of univariate data*.

One direction to extend this is to ask whether there is a measure of fit that the *median* minimizes. In DataSpace students can construct the sum of the absolute values of the residuals from the moveable line and discover the somewhat surprising fact that the median minimizes this sum. (If the number of points is even, there is a region between the two middle points for which the sum is a minimum.)

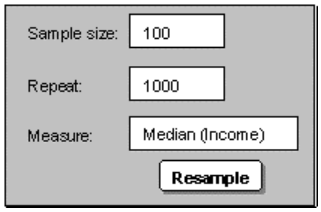


What about two dimensions? The plot at right shows shoe width versus person's height in centimeters with a moveable line and the squares of the residuals. Again, the moveable line is the model for the data and the sum of the squares measures the fit of that model with the data. By dragging the line around and observing the effect on the computed sum of squares of residuals, students experience, almost viscerally, the process of fitting a least squares regression line to bivariate data.

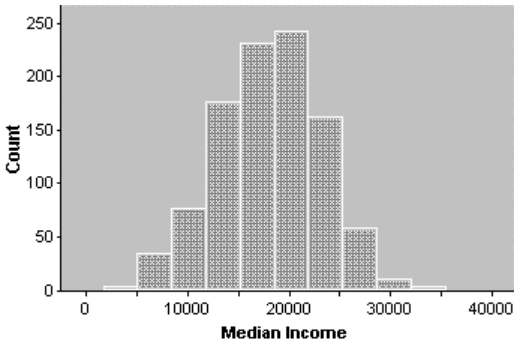
Simulation—Part of what makes inferential statistics infernally difficult for students is that it is mired in the hypothetical. Consider the phrase, "If I were to sample from the population again and again, …" But the fact is that we are generally *not* able to repeat the sampling process. With DataSpace, however, students make this hypothetical process real by using simulation.

Let's take a relatively simple example, estimating the median income for the population of a city, say San Francisco. (As part of the DataSpace project we are finding ways to make raw data from the 1990 US Census readily available.) We start with a 1% sample of all residents of San Francisco in 1990. In DataSpace this sample appears as a collection of people as shown at right. We agree that this *sample* of about 7,000 people is to be our *population*. One sample of 100 people from this group produces an estimate for the median income of the population. But how are our students to understand how and why we measure the uncertainty of that estimate?

In DataSpace, by connecting the collection to a *resampler*, we can peform the formerly hypothetical act of repeatedly sampling and computing the median income—or any other measure we might define. The resampler stores the results of each computation of the median in a new collection.



We make a histogram of the sample medians. A histogram of these values, as shown at right, lets us talk about and think about what confidence we have in our original measure.



We have only had the space here to hint at the pedagogic power that resampling techniques provide for teaching inferential statistics. With DataSpace we are striving to integrate resampling with other data-analysis tools because we believe that these techniques are extremely helpful for teaching statistics and because they are being more and more widely used in statistical practice.

Networked Collaboration

DataSpace is also fundamentally networked software. The "network"-ness is both local (within the classroom or school) and global (on the Internet). This has two immediate, interrelated consequences. First, it reduces the logistical nightmare of collecting and collating data—whether digesting an Internet data set so it can finally be imported, or collecting the whole school's survey results into a single file. Second, it lets

students share resources and collaborate, which is in itself good for learning and a good model of what real statisticians do.

In summer institutes for teachers we have tested prototypes of methods for collating survey data. In this setting we were able to demonstrate the efficacy of having students fill out survey forms on the computer, submit them to a central server, and go on to begin analysis of the data, all within a single class period.

Currently there are no useful standards for accessing raw data on the Web. The default standard of tab-delimited text with attribute names on the first line leaves unresolved the issue of how the data set and its individual attributes are documented. Query engines, such as that provided by the US Census Bureau, are particularly bad at producing data files that can be automatically imported into statistics software packages. We encourage statistics educators to join with government and industry to establish standards for storing and accessing data.

Our work with students sharing resources and results has barely begun but at least one surprising result has emerged: Ideas are data. Networked collaboration makes it possible to pool ideas instantly from a classroom full of small groups. Prioritization of ideas—voting on favorites—is also easy. Students scan through the list of submissions, choosing those ideas that they particularly like. Within a remarkably short time, a list of ideas sorted by how many students liked them, appears on the overhead LCD projector at the front of the classroom.

A Note on Classroom Testing of DataSpace

Classroom testing of the DataSpace software began in the fall of 1998. As this testing proceeds, and as the software becomes commercially available, we hope that our own group and other statistics educators will begin to assess the impact of this kind of computer learning environment on how students learn statistics.

---

[1] This material is based in part upon work supported by the National Science Foundation under award numbers III-9400091 and DMI-9660827. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[2] Jackiw, Nicholas, *The Geometer's Sketchpad*, Key Curriculum Press, 1991.