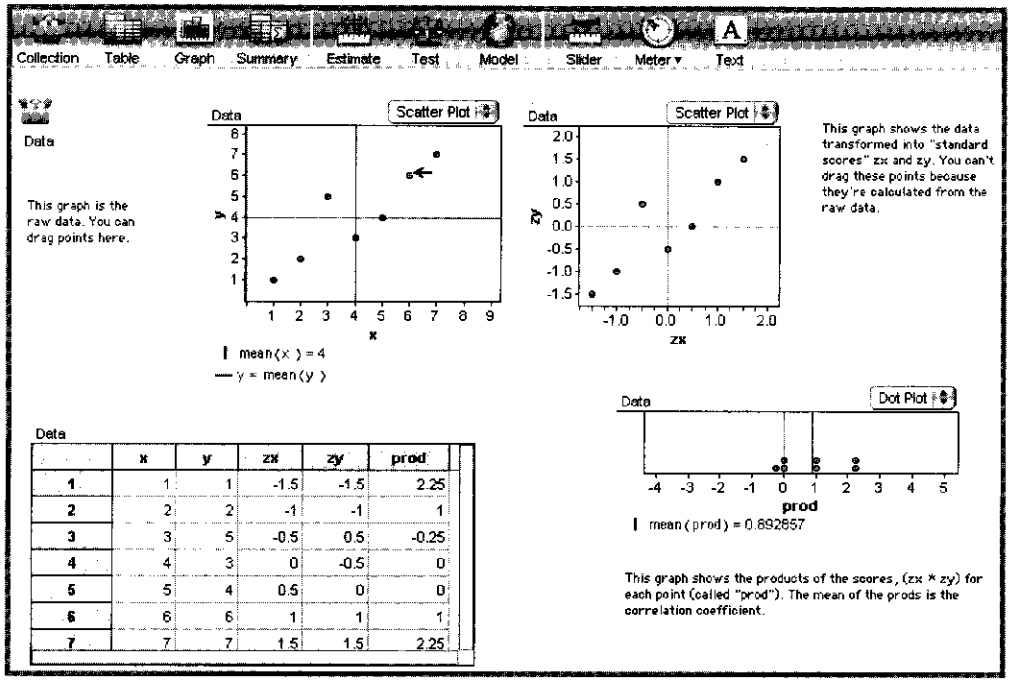
Demo 8: Devising the Correlation Coefficient

How the correlation coefficient measures what it does

Many students learn that a correlation coefficient near +1 or −1 means a perfect correlation, and they may even do exercises where they learn to recognize what a correlation of 0.8 looks like. But this demo is about how the correlation coefficient *works*: in particular, how it is that the product of the two coordinates—suitably transformed into standard scores—helps measure what we're after.



	x	y	zx	zy	prod
1	1	1	-1.5	-1.5	2.25
2	2	2	-1	-1	1
3	3	5	-0.5	0.5	-0.25
4	4	3	0	-0.5	0
5	5	4	0.5	0	0
6	6	6	1	1	1
7	7	7	1.5	1.5	2.25

What To Do

- Open **Correlation.ftm.** It should look like the illustration.

The graph in the upper left—where the pointer is in the illustration—shows the raw data, with lines showing the mean of **x** and the mean of **y.** To its right is a similar scatter plot that shows the "standard scores" **zx** and **zy.** Below that graph is a dot plot showing, for each point, the product of the two coordinates. That is, it shows **prod** where **prod = zx * zy.** The graph also shows the mean of **prod**—which is the correlation coefficient.

- Grab a point (such as the one indicated in the illustration) and drag it around the left-hand graph, watching how that affects the right-hand graph. (Ignore the bottom graph for now.)
- Drop that point near the intersection of the two lines—the intersection of the means of the two variables. Watch how the intersection moves as you approach.
- Move each of the points to a place near that intersection, watching how the mean moves and how the points behave in the other graph. Close is good enough. *Save one of the most extreme points for last.*
- Finally, move the extreme point to the intersection of means—and watch what happens in the other graph.

You should see that the standard scores—the *z*'s—mirror the data almost exactly, with a couple of crucial differences.

Questions

1 When you move a point, the others (on the **zx-zy** graph) first move in the opposite direction. Why? **Sol**

2 As you move the point even farther in one direction, what happens to the other points? **Sol**

3 How would you characterize the difference in shape between the top two graphs (that is, they look similar, but when you drag a point far to the right, they don't look as similar)? How do you account for that difference?

4 When you brought that last point in to the mean, what happened? Explain why.

Onward!

▷ Return the points to their original positions (or close to them) through repeated **Undo**, re-**Open**ing the file, and so forth.

▷ Now, move points around again, but this time look at the image of the moving point in the **prod** (lower-right) graph. Note where the point must be in order for its **prod** to be negative.

⇨ If you are leading a demo, this is a great place to involve the class in deciding, for example, where to move points.

▷ Move points so that they all have negative **prod**s. Notice where the points are.

▷ Now move the points so that they all have positive **prod**s. Again, notice the pattern.

▷ Make all the points line up perfectly with a positive slope. Notice the mean of the **prod**s (in the graph).

▷ Pick one point and move it along a line perpendicular to the line you have made. Observe what happens to the positions of other points on the **zx-zy** graph, its **prod**, and the mean of the **prod**s. Then return the point to its place in line.

▷ Again move the points in toward the center, one at a time. Watch the position of the extreme point—the one you're going to move last—in the **prod** graph.

More Questions

5 Where does a point have to be to have a positive **prod**? A negative **prod**?

6 When the points are all lined up, they still don't have the same **prod**s. Which points have big **prod**s? Which have little ones?

7 If you line up all the points perfectly with a *steep* positive slope, the original graph looks very different from the way it does if you line up all the points perfectly with a *shallow* positive slope. How do their **zx-zy** graphs differ? Why?

8 Why did the **prod** for the extreme point get more extreme as you dragged the other points to the center?

9 How did the correlation coefficient—the mean of the **prod**s—change as you dragged the points to the center? Why? **Sol**

Extension

▷ Move the points so that the mean of the **prod**s is zero. Do this in at least three radically different ways.

The Point

The correlation coefficient works because of the interplay of two big ideas.

- Looking at the products of the zx's and zy's is illuminating, because points in quadrants I and III contribute to the positive and II and IV to the negative parts of the correlation. Because the distances are all relative to the mean in each dimension, the sum of these products is a measure of how well all the points are lined up.
- Standard scores—the z's—scale the dimensions to show their spreads and distributions comparably, in a way that makes the original units and magnitudes irrelevant. Without using standard scores, you could still make numbers that measure how well the points are lined up, but you would not be able to compare those numbers to measures from other data sets.

Note: This transformation from x and y to zx and zy looks bizarre, but it's really exactly the same as any other translation and dilation in the plane. These are the same transformations that let you slide and stretch mathematical functions. The difference is that, here, the transformation depends dynamically on the data.

Challenges

10 When you line up the points, you can get a correlation coefficient—a mean of the **prod**s—equal to 1. Why can't you get the mean to be *larger* than 1? After all, some of the **prod**s are larger than 1; why can't you arrange it so that more of them are? Try to make a correlation coefficient larger than 1, and figure out qualitatively what happens to prevent you from succeeding. If you can, prove it with symbols.

11 Invent a correlation coefficient that's based on the median and IQR instead of on the mean and standard deviation. What possible values does it have? What are some advantages and disadvantages of using this measure of correlation?