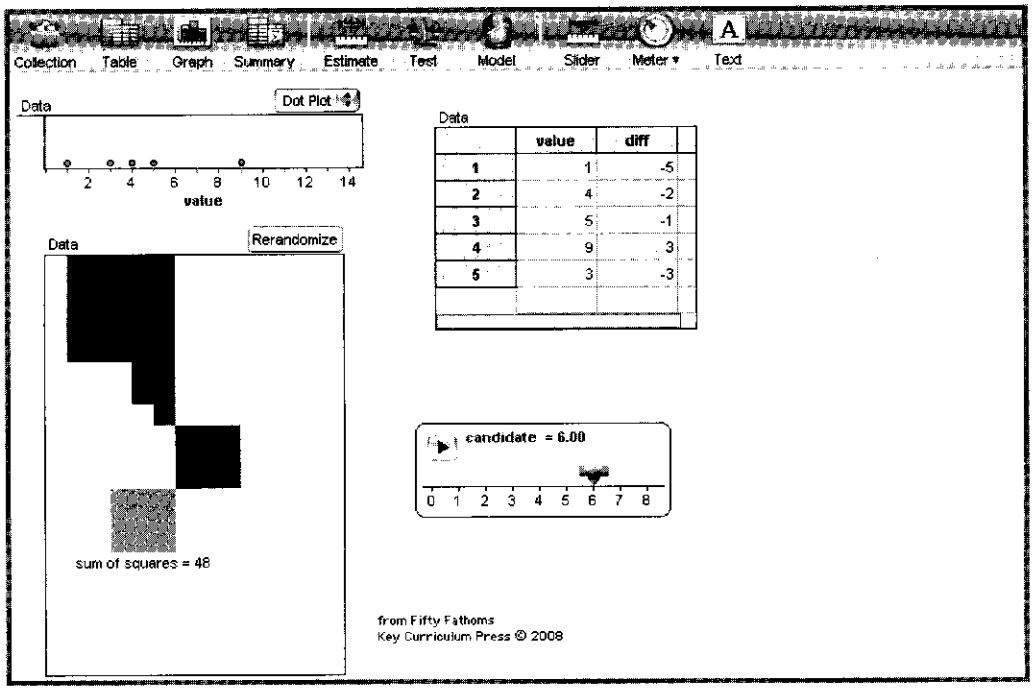
Demo 5: The Mean Is Least Squares, Too

Defining the mean as the place where the sum of squares of deviations is a minimum (just like the least-squares line) • The median, and what it minimizes

We have put this demo in the opening section because it is about the mean, and measures of center come early in this book. But this is a sophisticated demo. Demo 6, "Least-Squares Linear Regression," is more complex on the surface, but it's worth looking back at this one afterward to see the connections.



What To Do

- Open **Mean Is Least Squares.ftm**. It will look like the illustration.
  - Note: The **Rerandomize** button won't do anything because there are no random values.

Here we have five data points and a slider in the middle called **candidate**. That slider controls a number that we are proposing for a measure of center. We're essentially asking, "How good is this number—6.0 at the beginning—as a measure of center for these five data points?" (Lousy. But we're getting ahead of ourselves.)

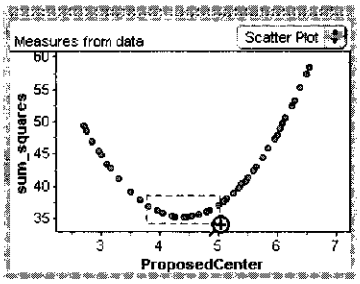
The collection itself (usually gold balls) this time shows colored squares. One vertical edge of each square represents the data value; the other is at the candidate value. You can also see that we have computed a quantity called **diff** for every data point, which is the difference between that point and the candidate (equal to the length of the side of each square).

- Drag data points in the dot plot (upper left) and see how the values—and the squares—change. Notice how the position of the dot corresponds to the vertical edge of one of the squares. Also, note how the sum of squares changes.
- Drag the slider, **candidate**, and see how things change. Try to move it so that the sum of squares (at the bottom of the stack of boxes) is a minimum.

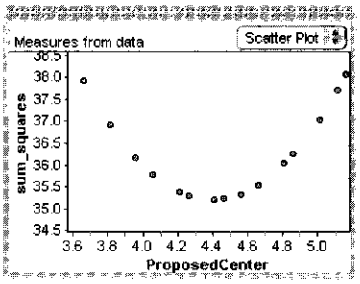
It would be so helpful to have Fathom collect all those sums of squares. Inside the **Data** collection, we have computed two measures. One, called **proposedCenter**, simply reports the value in the slider. But the other, **sum_squares**, is the sum of the squares of **diff**. That's what we're interested in looking at for different values of **candidate**.

(It's confusing that **proposedCenter** has a different name from **candidate.** But Fathom needs to distinguish between the attribute values and the slider.)

- Open **Mean Is Least Squares 2.ftm.** It will look like the previous file except that it has a measures collection and a graph showing how **sum_squares** depends on **proposedCenter.** We've already collected one value—the sum of squares is 48 when **proposedCenter** is 6.
- Drag the **candidate** slider. New points will appear on the measures graph, and the picture of the squares will update. Keep dragging until you have a clear minimum for **sum_squares** in the graph. The graph will look something like the one below.



- Holding down **Option** (Mac) or **Control** (Windows), drag a rectangle around the minimum (shown). When you let go, Fathom will zoom to that rectangle, and you'll see a clearer minimum:



- Add new points (by moving the slider to the relevant area) as you see fit. You may even want to zoom in to the slider to get finer values for **candidate.** Determine the value for the candidate that gives the smallest sum of squares of the differences.

This should be the mean, in this case, 4.4, and a sum of squares of a little over 35. That is, the mean minimizes the sum of squares of the differences—*the mean is a least-squares statistic.*

- Change some of the data values so that the mean is different.
- Delete all of the cases in the *measures* collection (not in the data collection!).
  - ⇨ One easy way is to select the graph, choose **Select All Cases** from the **Edit** menu and then choose **Delete Cases** from the **Edit** menu.
- Drag the slider again to produce the graph of sums of squares. See if that minimum matches the new mean.

Extension—What If We Don't Use Squares?

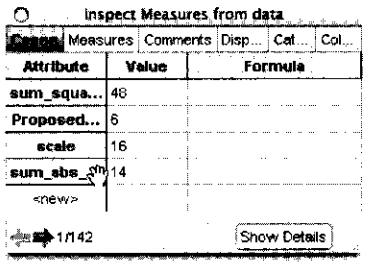
Adding squares of differences is a great way to get a reasonable measure of how good a candidate measure of center is. But it's not the only way to do it. In fact, when we ask people to devise their own measure for how good a center is, one of the most popular is to add the absolute distances of the points from the center. Speaking graphically, add the lengths of the line segments connecting the points to the center instead of the areas of the squares. Speaking symbolically, minimize

$$M=\sum|x-c| \text{ instead of } M=\sum(x-c)^2$$

where c is the candidate center.

Actually, to facilitate this extension, you've been collecting sums of absolute differences all along. All you need to do is graph them:

- ▷ Double-click the measures collection (the box of green balls) to open its inspector.
- ▷ Click the **Cases** tab to bring up the **Cases** panel.
- ▷ Drag the name **sum_abs_diff** (it will probably appear as **sum_abs_di...**) from the inspector (shown) to the vertical axis of the graph, replacing **sum_squares.**



- ▷ Drag the slider as necessary to fill in the graph. Where is the minimum value?

Extension Questions

1. What special value minimizes the sum of absolute differences? **Sol**
2. The stack of colored boxes in the collection on the left is shortest when this number is a minimum (not when the sum of squares is a minimum). Explain why.
3. Remove a case from the original collection: Select one of the colored boxes in the stack and choose **Delete Case** from the **Edit** menu. Now there are four instead of five. Then redo the graph of the measures (delete all cases in the measures collection, as described earlier and play with the slider). What do you notice about the graph? What's the minimum value? Explain.
4. Suppose that instead of adding squares or absolute differences, we minimized the (absolute) difference between the candidate and the *farthest* data point. What would the graph look like? What value would you get for a minimum? **Sol**
5. If you minimize the sum of absolute distances, or if you minimize the farthest distance (as in the previous question), your function of the candidate position is piecewise linear, unlike the smooth parabola we got for the sum of squares. What difference, if any, does that make in how useful these various minima are as measures of center?