Floyd on Slope Inference:

Katherine Cornett asks for suggestions for teaching about inference for the slope of a regression line, especially its standard error.

Floyd replies:

One of the reasons I think students struggle with this is that they don't actually see any variability in the slope of a regression line, since they only have one set of data. They somehow think that it's the only line that there could possibly be, rather than one among many hypothetical outcomes of a random experiment.

That is actually true of all the sample statistics in the course, but I think somehow the regression line slope is harder for the students to "get". I teach this topic, like so many others, using simulations.

I give the students a list of all the major league baseball players (any large list with bivariate data will do), and I have the students randomly sample five of them and plot their weights against their heights on their calculators. Then they are to construct the least-squares regression line (automatically, using their calculators) and make a note of the slope of the line. Then they are to repeat this process several times. While they're doing this, I'm putting a horizontal axis on the board labeled "regression line slope". Then after ten minutes or so, I have the students begin coming to the board and putting X's on the axis at locations corresponding to the slopes they found, building up a histogram of the slope.

Then we look at the histogram and discuss its properties. It typically looks pretty normal, which is something the kids notice. It is centered on the actual slope of the regression line for all the baseball players (I show them the population data in a scatterplot, along with its regression line), so the sample slope is an unbiased estimator of the population slope.

And then there's the standard deviation of the slope. The students can estimate it by looking at the histogram, but I point out that in practice, we don't get many samples, we only get one. In practice, our single sample serves multiple purposes: the slope of the regression line gives us an estimate of the slope of the population regression line, but the variability in the y-values above and below the line (along with their x-locations) also serves to estimate how much "wobble" there is in the regression line--that "wobble" is the variability in the slope, which is the standard deviation we see in the histogram.

Letting the data serve double-duty isn't new with regression lines. We do it even when we construct a simple CI estimating a population mean. The mean of the data serves as the center of the CI, and the sd of the data helps us estimate the SE of Xbar. Similarly, the data in a regression context both give us the line, but also estimates of the amount of uncertainty in the parameters. That's SE(b) for the slope.

--Floyd