Bad data illustrated

Unedited posts from archives of CSG-L (see INTROCSG.NET):

Date: Mon Jan 13, 1992 5:13 pm PST
Subject: Bad data

[From Bill Powers (920113.1200)]

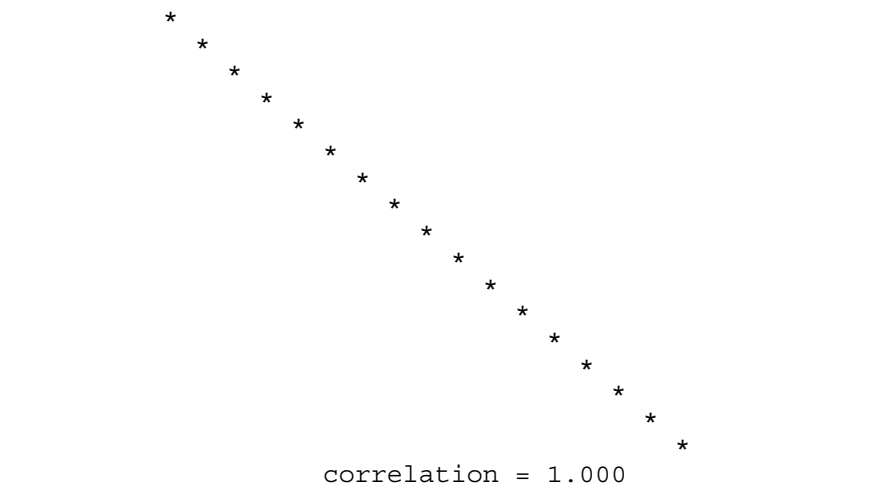
Still worrying the same bone -- what's wrong with statistical facts about individuals. I'm not bashing statistical studies of populations -- only the attempt to apply population statistics to individuals. I should mention in this context the modern classic on this subject by a CSG member, Philip J. Runkel: _Casting nets and testing specimens_; New York: Praeger (1990). A must-read for anyone who uses statistics in connection with human behavior.

My objection isn't esthetic or moral: it's that the predictions of individual behavior that come out of mass measurements are very poor, much worse than they need to be, mostly from lack of trying to meet higher standards for acceptance of facts. Today's offering concerns what predictions from bad data look like.

I wrote a little program that plots the function y = 2x + [a random variable]. The random variable is just the "random()" function from the C library, so it doesn't conform to Gaussian statistics, but the results are at least suggestive. What we're pretending here is that a dependent variable y has been postulated to be proportional to an independent variable x, and that this hypothesis is used to explain a collection of data points obtained by varying x and observing y. If there were a perfect linear relationship in the data, the points would plot as a straight line. After generating an array of 24 pairs of data points, we calculate the correlation coefficient between x and y. The question then is, how well does the regression equation, y = 2*x, predict the value of y given the value of x?

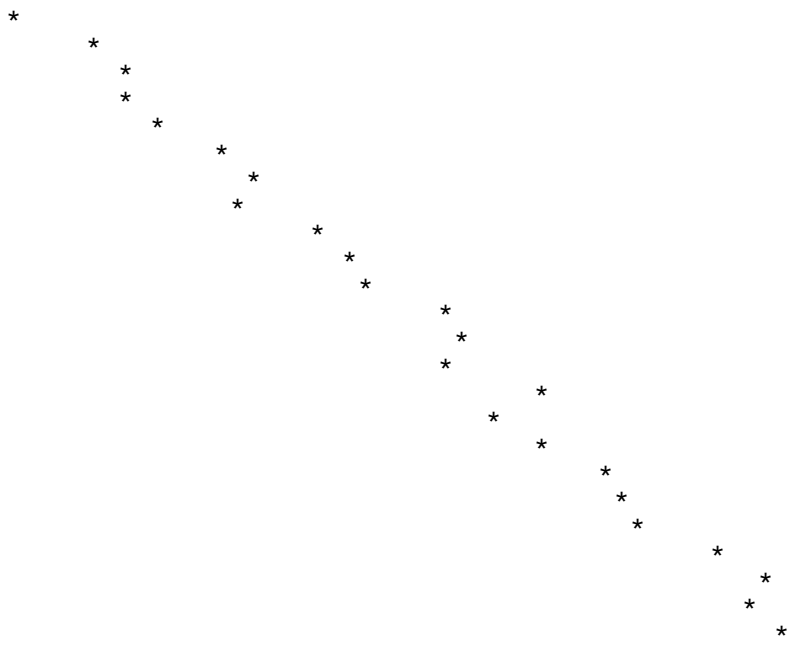
In the plots below, x runs from top to bottom and y runs from left to right.

Here is the plot of y vs x when there is no random noise added to the measure of y:



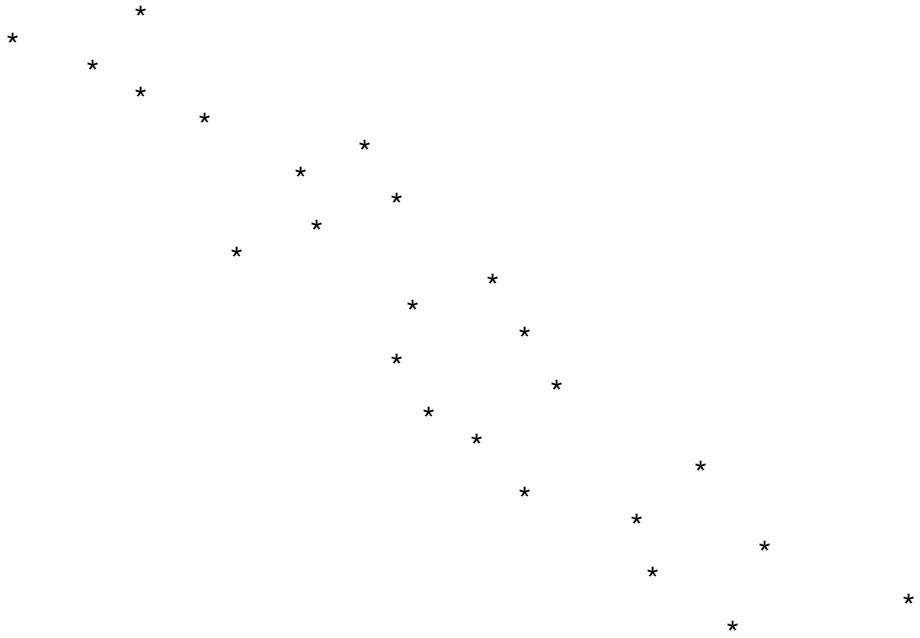
Obviously, given x you can predict y exactly. There is no scatter.

Here is what the data look like when enough noise is added to bring the correlation down to the level we get in easy tracking experiments:



Correlation = 0.995

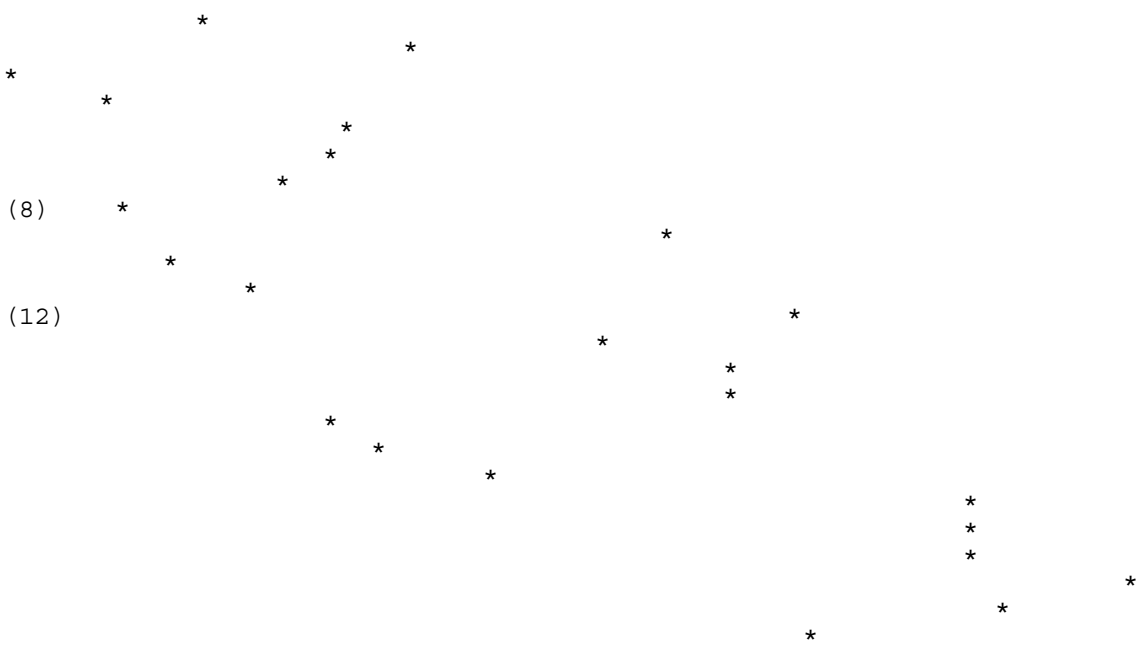
When handle sensitivity gets too high or disturbances get large, the correlation drops to the low 90s, something like this:



Correlation = 0.928

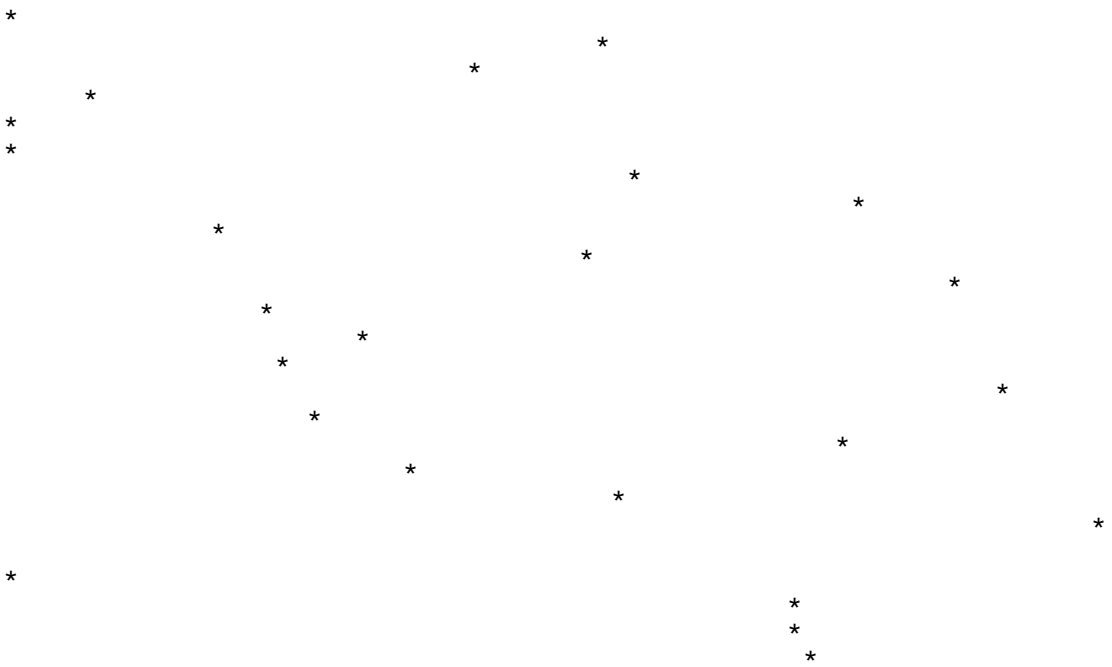
In most statistical studies of relationships between dependent and independent variables, a correlation of 0.8 would be considered very high.

Here is what the data would look like:



Correlation = 0.798

Even a correlation of 0.6 is considered rather good:



Correlation = 0.620

As Gary Cziko has reported, there have been published studies in which relationships with correlations of 0.2 have appeared.

Here is that degree of correlation:

Correlation = 0.201 (the points on the left actually went somewhat to the left of zero)

An interesting fact came up while I was generating these plots. When the argument of the random function is set to produce a correlation of 0.6, and the plot is generated over and over, the result can be any correlation between 0.3 and 0.8 on repeated trials, as different sets of 24 random numbers are generated. The implication is that with only 24 subjects, one can't say what the meaning of a given correlation is without re-doing the study many times. The first correlation obtained is very unlikely to be at the center of the spread of correlations. How many times do typical researchers replicate their studies, to find where the center of the range is? I suspect that the mean number of replications is close to 0.

Suppose that a person is exposed to 12 units of the independent variable x (top to bottom, halfway down). You want to use this score to predict that person's score on a test of the dependent variable y (left to right). Looking at the above plots, at what level of correlation would you begin to take the prediction of y seriously for that person? I would say that at r = 0.8, the prediction is too bad to use: clearly, the error in prediction would be something like 50% of the y-score. I wouldn't be much interested unless unless the correlations were up into the 0.90s.

Suppose that you were comparing two people, one with an x-score of 8 and the other with an x-score of 12. This would be like using one questionnaire to determine the independent variable, and using some other measure of the dependent variable. That's a difference of 4 points around the average of 10, or a 40% change in x-score. I've labeled the 8th and 12th lines in the plot for a correlation of 0.8. Clearly you would get the right comparison and then some. But suppose you move them both up one notch, or two, or three. Your prediction could differ from the actual difference in y scores by a large amount -- it could easily be backward.

Again, I don't think that any correlation lower then the 0.90s would be scientifically usable. And you don't get results that you could call *measurements* until you're up around 0.95 or better.

When you look at the plot for a correlation of 0.6, it's easy to see the trend. Clearly there's something going on here that you can see with the naked eye, despite the huge scatter. An effect! It's easy to overlook the fact that in order to see this "effect," you have to look at ALL the data points. You

don't get this impression from looking at just a few of the points (put your hands over the plot so you can just see the center part). This "trend" you see is a property of the whole plot. The individual measurements don't "trend." Each point is where it is. The trend line, y = 2x, is far above many points and far below most of the rest. The distance from the trend line for each point shows you how badly the trend line misrepresents each point.

When you use the trend line to predict differences between people, the picture gets even worse. By drawing a line between various pairs of points, you can get slopes ranging from highly positive to highly negative. But the trend line predicts that the slopes should all be the same as the slope of the trend line. You have to get high into the 0.90s before comparisons mean anything at all.

There's another way to look at this. Somewhere around the 0.80s, the scatter becomes small enough that you could divide the y scores into a high group and a low group. You could then say that if the x score is less than, say, 6 or greater than, say, 18, it will predict that an individual point is in the low group or the high group. What has happened here is that the resolution of the "theory" y = 2x has become just great enough to treat the measurements as binary data: 0 or 1. We can pretty well tell the difference among 0,0 0,1 1,0 and 1,1. As the correlation rises above 0.8, the coarseness of the meaningful numerical measures falls: we begin to make out details. And when the correlation is in the upper 90s, we begin to get something resembling a continuous measurement scale.

When the resolution is too low, most of the data points are useless; it takes an extreme of the independent variable to predict that the dependent variable will be in the high group or the low group. In this case, the useful N is not the total number of subjects or points. It is a much smaller number, only the points indicating extremes of both x and y. Below correlations of 0.8, most of the points near the middle are useless. Even at 0.8, all we have is a crude measure that could easily be confounded by any slight effect from a common cause.

A true science needs continuous measurement scales so that theories about the forms of relationships can be tested. This means that correlations have to be somewhere in the high nineties. True measurements, with normal measurement errors, require correlations of 0.99 upward. If this were universally understood among scientists, two things would happen. The first is that most statistical studies would end up in the wastebasket. The second is that the good studies would be done again and again, with successive refinements to reduce the scatter, until something of actual importance and usefulness was found.

Best, Bill P.