OTHER THINGS EQUAL

Donald N. McCloskey
University of Iowa

The Bankruptcy of Statistical Significance

Kenneth Arrow visited Iowa the other month. My colleagues and I ate dinner with him at the local Italian place and went to his talk afterward, on why public higher education should not be free. The encounter reminded me why Arrow is widely considered the best American economist of his generation. He got an MA in mathematics as a youth, but nonetheless the man somehow learned to be an *echt* economist, no mechanic. One sign was that his main argument against subsidies for research though subsidies for undergraduate education is accounting: namely, after all only 1 percent of our graduates go on to graduate school. The other was that he was willing to talk economics right through the meal.

I brought the conversation round to statistical significance -- that standby of gracious dinner conversation among economists -- and reminded him that in 1959 he wrote a paper in an obscure *festschrift* (for Harold Hotelling, widely considered the best American economist of *his* generation) saying that statistical significance is useless. Arrow corrected me: he had not said that it is useless, merely grossly unbalanced if one does not speak also of the *power* of the test. But, I replied, we never do speak of power. "Yes," said Arrow, "I agree. Statistical significance in its usual form is indefensible."

Then he said something surprising. I was the only person, he claimed, to pick up on his 1959 article, in a little squib in the *AER* in 1985 called "The Loss Function Has Been Mislaid" (one of the many titles after Bob Gordon's original classic of empirical economics in the 1969 *AER*, "$45 Billion of U. S. Private Investment Has Been Mislaid"). What is so surprising? This: the best economist of his generation says, in effect, "Folks: your main method for running empirical work is *indefensible*," proving so beyond rational objection, and yet practically no one pays attention.

Maybe it was the obscurity of the outlet, but I don't think so. The same message, with nearly the same quality of messenger, making one or another devastating criticism of statistical significance, has been delivered again and again and again in places anything but obscure: Edward Leamer's "Let's Take the Con Out of Econometrics" in the 1982 *AER* is one instance; Tom Mayer's "Selecting Economic Hypotheses by Goodness of Fit" in the 1975 *Economic Journal* is another; Gordon Tullock's one page blast in the *Journal of the American Statistical Association* back in 1959 is another. And when you start looking into it you find people making the same crushing points over and over again across the sciences. A famous psychologist, who gloried in the name of Edwin Boring, made the central point in the *Psychological Bulletin* of 1919. A great statistician, William Kruskal, reminded us of it in his article on "Statistical Significance" in the old *International Encyclopedia of the Social Sciences* [1968], part of which was revamped as the *International Encyclopedia of Statistics* [1978]. Another great statistician, John Tukey, in his "Sunset Salvo" in the 1986 *American Statistician*, likewise attacked the gross misuse of significance levels. For that matter, the original article of 1933 on which modern statistics is built, by Neyman and Pearson in the *Philosophical Transaction of the Royal Society*, Part A, makes the point with an example of a criminal

Eastern Economic Journal, Vol. 18, No. 3, Summer 1992

conviction [1933, 296]. Among economists, Ed Feige, Zvi Griliches, Knox Lovell, Frank Denton, and many other have demolished statistical significance in public for all to see.

Yet no one sees. Most recently the man from whose book I learned econometrics back in the 1960s, Art Goldberger, has again made clear that something is deeply wrong with the economic use of statistical significance, on page 240 of his excellent new textbook. But I'll be amazed if anything changes. Statistical significance is bankrupt, its assets valueless, its liabilities growing by the minute. But the creditors have simply decided to keep on pretending that checks signed "S. Significance" are worth banking on. The result has been a scientific inflation, a regular bubble.

Step back for a minute and recall what a "significant" coefficient means. It means that the *sampling* problem has been solved, or at any rate solved well enough to satisfy conventional standards. That is all it means, and all its mathematics justifies. In other words, *the sample is large enough to assure that if you took another sample it would give roughly the same result.* The sampling variance, which is the population's variance divided by the square root of the sample size, has been driven down to some nice, low figure. As John Venn put it in 1888, at a time when our procedures were a mere twinkle in the statistician's eye, the coefficient (or the mean or the difference between two means or the estimated variance of the R-squared or whatever other statistic we are examining) would probably be "permanent". We would come up with the same estimate again.

But a permanent coefficient is not necessarily an important coefficient. *That's the main point.* Forget all your cynical, if true, jokes about trolling through the data for the significant coefficients. Sure, statistical significance doesn't mean what it claims to mean if, as one naive student admitted to his thesis committee, you have run fully 200 different specifications of the same economic idea. But set that point aside. The main point -- which would remain true of the most virginal classical regression on an absolutely fresh cross-section, a literal and unbiased sample from a well-behaved universe, with perfect specification, complete agreement on the Type I error, full treatment of the power function, and an honesty in handling the data to the standard of Mother Teresa and George Washington combined -- is that a coefficient of, say, 1.3567 on X is not *scientifically* significant unless it is interestingly big or small or close to 1.00 or different from zero or in the neighborhood of 1.8923 or whatever by *scientific* standards. The coefficient of 1.3567 might be statistically significantly different from, say, zero at the .0000000000001 level of significance (and would in fact be so if the sample size were large enough). Yet its permanence, speaking of sampling variability, at just about exactly 1.357 does not make it important. If the question asked by putting X in the regression is scientifically unimportant, what does it matter if the answer is permanent? "Statistically significant" does not mean "substantively significant". The two significances have nothing to do with each other.

What matter, to use a technical term, is *oomph*. Oomph is what we seek. A variable has oomph when its coefficient is large, its variance high, and its character exogenous, all decided by quantitative standard in the scientific conversation. A small coefficient on an endogenous variable that does not move around can be statistically significant, but it is not worth remembering. Oomph is what we mean when we talk about money being "important" for explaining the price level or about capital being "important" for explaining income per person. A record corn crop in Iowa (yes, it was, thanks) certainly does raise average national income, however hard it might be to discern in the national noise, but has little oomph because the coefficient on the Iowa-corn-crop regressor is doubtless

low. Likewise, the existence of oxygen in the atmosphere certainly does affect combustion, but it does not vary enough to give it oomph in an explanation of why the house burned down. The stock of money in the hands of Iowa Citians certainly does determine their expenditures, but because it is entirely endogenous it has no oomph.

The best way to see the point is to suppose that you really know what some coefficient is. For sure. God has told you, with no nonsense about confidence intervals; sampling error is zero. The *t*-statistic is infinite, which should satisfy the most knuckle-headed reviewer of your paper. Well, then: Has the variable got oomph? Go ahead. Think about it.

Time's up. The answer is, *You don't yet know.* To find out you have to ask and answer different questions, having nothing to do with statistical significance, such as whether the coefficient is large, (how large? Large enough to matter in some conversation of scholars or policy makers), or whether the variable could vary enough to produce effects you consider important. For most scientific or policy questions the answer that across successive samples with a nice, random character the coefficient would be permanent in repeated samples is only mildly interesting.

So what? Here's what. Almost all econometric fittings have to be done over again. All of them. All the statistical work that has dropped and added variables by statistical significance needs to be redone. None of the econometrics that decided whether variable X is "important" by using statistical significance has been correct, for all these years. It's good news for assistant professors: all the work of your elders has been wasted, which leaves you with a brilliant career ahead redoing what they did wrong.

Some will say: but only bad economists do such bad things. Ho, ho. Look at the latest issues of the *AER* or any other journal of economics. I will present a $100 check to anyone who can show to the satisfaction of a panel of world-class statisticians (I get to choose them, but trust me) that more than a small fraction (a quarter, say, to be sure of my bet) of the empirical papers do anything but grossly confuse statistical with scientific significance.

Want to make yourself unpopular with your colleagues? Xerox this piece and put it in their mailboxes. Worse, go read the literature against statistical significance and start asking your colleagues if they know what they are doing. And here's a question to ponder. Eminent statisticians and many econometricians declare statistical significance to be bankrupt. Yet scientific practice does not change at all. The textbooks go on miseducating the students. What's going on? Moral failure in the profession? Carreerism gone mad?

I dunno. You go figure. But when figuring don't use statistical significance.

Other Things Equal, a column by Donald McCloskey, appears regularly in this *Journal.*