The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives
Stephen T. Ziliak and Deirdre N. McCloskey
Ann Arbor, Mich.: University of Michigan Press, 2008, 352 pp.

How do many scientific disciplines estimate and report results? Practitioners estimate regression models or conduct difference-of-means tests through experiments. And they report which results are significant and which are not (i.e., different from zero with 95 percent confidence). In this important book, Ziliak and McCloskey have three objectives: to remind us that such research may be mindless, unscientific, and costly; to explicate the intellectual history of significance testing and the struggles among those professors who developed sampling and statistical testing; and to illustrate the correct way to conduct research and praise those few who report their research properly.

The Costs and Benefits of Significance Testing

First, a little review of significance testing. The central question in research is what is the effect of some variable of interest on an outcome. In medicine, for example, we want to know the effect of a drug on illness. In economics, we care about the effect of prices on consumption or work choices. To assess those effects, we rarely have data on populations. Instead we have data on only hundreds or sometimes thousands of people.

Researchers must estimate the likelihood that the results from the sample represent the results if the population were studied. The answer depends on the size of the sample and the signal-to-noise ratio in the sample. The smaller the sample and the smaller the signal-to-noise ratio, the lower the likelihood that the sample result is the population result. Said differently, small sample sizes and noisy data increase the variety of possible population results that are logically possible given a particular sample result. In such small, noisy samples, it becomes more likely that observed effects are the result of chance rather than systematic factors.

And then there is the question that actually has no scientific answer: How confident should we be that a result is not the result of chance? This book chronicles the development of the convention that 95 percent likely is likely enough and then the degradation of

that convention into what the authors view as its fatally flawed shrivelled version: unless a sample result is different from the result of zero with 95 percent confidence, you have no result at all.

What is so odd about the role of statistical significance is how out of character it is for economics. Normally economists preach to other disciplines to think continuously rather than dichotomously: how much more or how much less rather than right or wrong, yes or no. Confidence in results is a continuous function: 96 percent confidence is more than 89 percent confidence. But in a world in which confidence has degraded into significance tests, 89 percent is unacceptable (no result) and 96 percent is just fine.

To be sure, benefits exist from the adoption of a rule that keeps so-called false-positive or Type I errors to less than 5 percent. Under such a rule, a researcher must be 95 percent confident that a result could not have arisen by chance even if the *unobserved* underlying truth is no effect or no result. Under such a significance rule, for example, pharmaceuticals are kept off the market if there is 6 percent or more possibility that a positive health effect could have arisen even though the drug actually has no positive effect at all.

But there are costs to such a rule as well. A central point of the book is to remind us of the costs. The more one attempts to reduce the possibility of false-positive statistical errors, the larger the possibility of false-negative or Type II errors. False negative errors occur when the underlying unobserved truth is different from zero effect. Such a result cannot be differentiated from a false positive result with 95 percent confidence and so is declared to be no result even though its effect is real. For example, false negative errors keep drugs off the market that would provide health benefits. How odd it is for economics to have developed a convention about the reporting of results that focuses only on the benefits of Type I error prevention rather than balances the costs and benefits of Type I and II errors.

Some Examples of the Costs

While I agree with their arguments that Type II errors have costs and that good research should discuss the costs and benefits of both Type I and II errors, the subtitle of their book (*How the Standard Error Costs Us Jobs, Justice, and Lives*) and some of the examples in their book lack the nuance and care that is the basis of their criticism of others' research. In chapter 16, the authors describe an article from

the *New England Journal of Medicine* that describes the possibility of false negative errors in drug trials with small sample sizes. The article correctly states that many of the therapies categorized as "no different from control" *could have had positive effects*. "Concern for the probability of missing an important therapeutic improvement because of small sample sizes deserves more attention in the planning of clinical trials" (p.179). But later in their discussion of the *New England Journal* article, Ziliak and McCloskey overstate the same concern:

> Yet they found that 70 percent of the alleged "negative" trials were prematurely stopped, missing an opportunity to reduce the mortality of their patients by up to 50 percent. Of the patients who were prescribed sugar pills or otherwise dismissed, in other words, about 30 percent died unnecessarily. In one typical article, the authors in fact missed at $\alpha = 0.05$ a 25 percent reduction in mortality with probability about 0.77 and, at the same level of Type I error, a 50 percent reduction with probability about 0.42 [p.180].

Notice the distinction between the possibility and certainty of Type II errors has disappeared. In order to make precise claims about the probability of a particular Type II error occurring, one has to assume the actual magnitude of the effect under consideration. But the actual magnitude is never really observed. Statements like those in the original medical journal article that say therapies *could have had positive effects* are accurate. Statements that say 30 percent died unnecessarily are overstated and misleading because the underlying truth necessary for such a calculation is not observed.

Another case example used by Ziliak and McCloskey to illustrate their arguments also seems to be off the mark (p. 94). A cost of too much attention to statistical significance is the lack of attention to the economic significance of the coefficients in a model. Again I agree with their general argument. But they illustrate their claim with an example from Milton Friedman's experience during World War II. He used regression analysis to estimate the effect of metallic composition on the fatigue of blades in turbo superchargers in airplane engines. Then best practice was breakage after 20 hours of use. Friedman predicted 200 hours of use from a change in alloy composition. But when metallurgists actually tested such alloys they broke after only 2 or 3 hours of use.

Ziliak and McCloskey argue that the case illustrates that statistical significance is not substantive significance. Now it is certainly possible that 200 hours was the estimated prediction given the coefficients for the effect of alloy composition on time before failure and that this prediction was different from zero with 95 percent confidence. And it is possible that 2 hours was within a 95 percent confidence interval centered on the predicted result of 200 hours. And both those possibilities combined with the experimental result of alloy failure in 2 hours would certainly be a good illustration of actual outcomes differing by two orders of magnitude from the predicted outcome—even though the predicted outcome was statistically significant. But the reader is not given enough information to make that determination.

But even if the reader were given enough information, the conclusion would seem to go against their general argument that exclusive concern with only Type I errors is the problem. That is, even though the predicted mean of 200 hours before failure was much better than current practice (10 to 20 hours), outcomes worse than current practice but greater than zero were within the 95 percent confidence interval. If one wanted to ensure that the possibility of outcomes worse than the status quo would be extremely low—let's say 1 percent—then that would necessitate the use of even more stringent Type I error prevention rules, which is the opposite of the thrust of their book.

I agree with their argument that researchers should describe their results carefully and completely. But their repetitive strident attacks on most economists and those in other disciplines for their failure to describe their results with sufficient precision and with all required caveats leave them open to the same criticism that they levy against others. And that is unfortunate because some of their examples (two of which I have described here) seem to be unclear, at best, and maybe wrong. And that gives sceptical readers an easy rationale for ignoring their central message, which I think is correct.

Why Does Economics Use Significance?

The failure of most economists to follow the example set by Ziliak and McCloskey is puzzling not only because economics is naturally continuous rather than dichotomous in its thinking but also because economics has been quite receptive to other methodological correc-

tions. The rational expectations revolution in macroeconomics, the concern for selection bias in all data that are natural rather than random, and the increased scrutiny of all time series regression results (unit root tests) all originated as challenges to orthodoxy but quickly became orthodox. Thus, the resistance to the arguments of Ziliak and McCloskey seems to me to be the exception rather than the rule in the attitudes of economists toward methodological improvement.

And why has economics gone down the wrong path on significance but progressed in those other areas? Ziliak and McCloskey do not discuss the lack of progress in significance testing relative to methodological progress in other areas such as the three examples I mentioned. Instead they relentlessly document the lack of progress and then narrate a person-centered explanation that blames Ronald Fisher and his followers. As with the use of the two cases I described earlier (the *New England Journal* and Friedman cases), the use of this type of historical explanation gives readers who do not accept this line of explanation a rationale for ignoring the methodological arguments of Ziliak and McCloskey.

Should Science Govern Choices?

What message does the book have for libertarians? Ziliak and McCloskey hint at the answer in their discussion of a methodological survey article in psychology. In the article the author calculated that the probability of mistakenly rejecting a treatment that actually had a large effect was 17 percent. Ziliak and McCloskey comment that "if you were dying of cancer, you might not view a 17 percent chance of needlessly dying as 'satisfactory'. . . it would seem that a better formulation in medicine is that it 'must be left to the patient, friends, and family'" (pp. 136–37).

Many health and safety decisions are delegated to bureaucracies, like the FDA, that allegedly use scientific methods to decide what products and practices to allow on the market. In fact values enter into such decisions in three ways. First, scientists have to decide how large the clinical trial sample sizes should be because that, in turn, dictates whether small effects can be differentiated from zero effect. Second, they have to either accept conventional significance tests or propose alternatives, and this choice dictates whether Type I or II errors are more likely and thus implicitly less costly. Third, given the findings of clinical trials, scientists and doctors and eventually the

FDA itself, in turn, vote on whether the benefits are worth the costs, which is obviously an economic rather than strictly scientific decision. In a more libertarian world, government or preferably multiple private entities would gather and disseminate information in a manner informed by the arguments of Ziliak and McCloskey but then let the public decide what to do with it.

Peter Van Doren
Cato Institute