STATISTICS AND SOCIETY

Signifying Little

Theodore M. Porter

The Cult of Statistical Significance
How the Standard Error Costs Us Jobs, Justice, and Lives

by Stephen T. Ziliak and Deirdre N. McCloskey

University of Michigan Press, Ann Arbor, 2008. 348 pp. $75, £48.95. ISBN 9780472070077. Paper, $24.95, £16.50. ISBN 9780472050079. Economics, Cognition, and Society.

"Repent," we are commanded in the conclusion to *The Cult of Statistical Significance*. "Sell all your goods and come with us." From Karl Pearson's "Saint Biometrika" at the beginning of the 20th century to the gospel according to Ronald A. Fisher, the cool, technical reason of statistics has also inspired messianic anticipations, and Stephen Ziliak and Deirdre McCloskey are only partly joking when they propose that we can enter the promised land of effective science only by rejecting Fisher's methods. The "standard error" of their title—the one that costs us jobs, justice, and lives—is the confusion of statistical with substantive significance. The authors do not claim much originality in recognizing this as an error; they list and discuss a distinguished roster of predecessors in statistics, philosophy, and the sciences who have called attention to it. And yet somehow the error persists, across a wide range of disciplines including some—such as pharmaceutical regulation, econometrics, and education studies—that feed directly into policy. The book was written to shake us out of our lazy habit of treating significance levels as an almost automatic criterion of scientific and practical worth.

If not Fisherian significance, what should be the Holy Grail of statistics? Ziliak and McCloskey (economists at, respectively, Roosevelt University and the University of Illinois at Chicago) answer: "Oomph." We should identify quantities that matter and measure them, not merely determine whether they can be distinguished from the null (meaning no effect) at some predetermined likelihood level. The validity of this point I take to be virtually self-evident. Yet statistical tests that ignore quantity remain pervasive, as the authors demonstrate through quantitative analyses of the contents of some very prestigious journals of economics, psychology, and medicine.

Of course, effective measurement requires in addition evidence as to the accuracy of the measure. The authors are not much interested in this aspect of the problem (the estimation of statistical error), and perhaps they are too willing to advise acting on the basis of inferred causal relations that, if assessed in a Fisherian way, admit serious doubt as to their very existence. At a minimum, it will normally be advisable to continue investigating when the evidence suggests, even tentatively, that a new drug causes suicides or a social program to help people find work saves three times what it costs. The book's inattention to sampling error may perhaps be forgiven as a corrective to the usual preoccupation with it.

Fisher and his disciples, I would agree, have a lot to answer for. But Ziliak and McCloskey also take some cheap shots, blaming Fisher for transgressions by medical and social researchers that he did not endorse and would not have countenanced. The most blatant of these is the supposition that a failure to demonstrate statistical significance licenses the assumption that an effect or causal relation does not exist. Ziliak and McCloskey show how often this move is made by economists running regressions and medical researchers analyzing experiments. But we should consider the following example, the main support for the assertion in the book's title that confusion about significance "costs us … lives." In a clinical trial of Vioxx in 2000, five experimental patients suffered heart attacks, compared to only one in the control group. The different did not rise to statistical significance at the 5% level, and on that basis the researchers declared there was no danger. Is this dubious reasoning to be blamed on the Fisherian statisticians? Not really. The sins of pharmaceutical trials are legion, as is now amply documented, and in this case the work was done and the paper written by Merck, which subsequently recruited phantom authors at universities willing to attach their names to the publication. Ziliak and McCloskey point out that Merck had already suppressed other data implying the riskiness of Vioxx. Most of the really harmful abuses chronicled in the rather overblown first 40 pages of this book involve statistical maneuvers that are illegitimate even by the standards of Fisher's program. We can better explain the Vioxx episode in terms of a corrupt research program than of flawed statistical tools.

The defects of Fisherian theory, it is clear, were made much worse in the practices of social and medical researchers as these were institutionalized from about 1930 to 1960. Most of the problems arose from an effort to set up inferential statistics as a set of recipes that could obviate any need for good sense of judgment. Ziliak and McCloskey argue that anxieties about subjectivity in the social and medical sciences encouraged reliance on such "magic pills." In my view, this quest for mechanized objectivity is typical of the modern adaptation of science to the state, which generally places scientists in a subordinate position when public decisions are at issue and demands of them, above all, rigorous impersonality and value-neutrality.

The book connects its statistical claims with a dubious morality tale. Even if William Sealy Gosset (who studied statistics in Karl Pearson's laboratory and worked out the basics of the *t* test) was as saintly as he is portrayed here, we might doubt the strongly implied link between his moral goodness and the correctness of his statistical program. On the chief point in question, the value of measuring effects rather than merely controlling error, he does not appear to me to be very different from that famously difficult personality and committed eugenicist—and effective founder of the 20th-century mathematical field of statistics—Pearson. The authors' hagiography of Gosset is counterpointed by a demonology of Fisher. Fisher's unattractive eugenic politics, refusal to give credit to colleagues and predecessors, and personal cruelty may also be genuine, but can these explain the misdirection of his statistical program? A shift in statistical practice from detecting effects to measuring them and assessing their consequences would be good for science, but it may not hasten the millennium.

"Student" of the *t* test. William Sealy Gosset.

CREDIT: COURTESY THE GALTON LABORATORY, UNIVERSITY COLLEGE LONDON

The reviewer is at the Department of History, 6265 Bunche Hall, Box 951473, University of California, Los Angeles, Los Angeles, CA 90095–1473, USA. E-mail: tporter@history.ucla.edu

10.1126/science.1160305