From *SIAM News*, Volume 41, Number 9, November 2008

Aboard the Statistical Significance Testing Bandwagon

The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. *By Stephen Ziliak and Deirdre N. McCloskey, University of Michigan Press, Ann Arbor, 2008, 352 pages, $24.95.*

BOOK REVIEW

By James Case

Statistical significance, as determined by *t*-tests, *F*-tests, and the like, is not to be confused with worldly significance. Conflation of the two, according to the authors of the book under review, has been responsible for a veritable river of bad science. No few others agree. In fact, the ink was barely dry on the tables William Sealy Gosset (a.k.a. "Student") compiled—in or about 1908—for the performance of his *t*-test when he began to warn of the poor decisions to which undue reliance on such tests can lead. Even so, his warnings came too late! R.A. Fisher had already embarked on his ultimately successful campaign to popularize statistical significance testing, despite cautionary signals from his friend Gosset. Even the few who realized that Gosset, not Fisher, was the natural father of statistical significance testing boarded the Fisher bandwagon.

Born in Canterbury in 1876, Gosset read mathematics and chemistry at New College, Oxford, before joining the Dublin brewery of Arthur Guinness & Son in 1899. There he applied his technical knowledge to problems arising both on the farm—where the culture of barley was of continuing concern—and in the brewery. His most famous papers appeared in 1908, shortly after his return from a leave of absence spent in Karl Pearson's biometric laboratory. Long accustomed to dealing with large samples, Pearson helped Gosset with his mathematics, but failed to share his abiding interest in small samples. Fisher, on the other hand, was quick to see the value of Gosset's methods, and received with his pre-publication copy of "Student's Tables" a note addressed to "the only man that's ever likely to use them."

The quantity tabulated was $z = t / \sqrt{n-1}$. The switch from z to t was Fisher's doing, as the latter was better attuned to his focus on degrees of freedom. Gosset was obliged by Guinness policy to publish under a pseudonym, papers by previous company researchers having been found to reveal trade secrets. He remained in Dublin until 1935, when he transferred to a new Guinness brewery outside London. He died of a heart attack two years later, at the age of 61. His collected works were published in 1942. Friends described Gosset as an unfailingly modest man who regularly deflected praise for his work, remarking that "Fisher would have discovered it all anyway."

The weakness critics see in tests of statistical significance is that they ignore the relative magnitudes of the consequences of false positives and false negatives. The authors liken the use of such tests to a decision about crossing a busy street. Whereas the attendant risk may be worth running to prevent a toddler from wandering out into traffic, it seems unjustified by the desire to obtain a second packet of mustard for one's hot dog. For them, "size matters." Loss functions, though ignored by *t*-tests and *F*-tests, are an important part of statistical decision making.

The authors often use the term "oomph" as a synonym for the size of the effect in question. A test showing that a certain medicine is rarely effective, yet saves lives when it works, has oomph. One showing that such a medicine saves only a few cents on the rare occasions when it works does not. They cite Jeffries's *d*, Wald's "loss function," Savage's "admissibility," Wald and Savage's "minimax," and Neyman-Pearson's "decision" as useful measures of oomph.

Although the book is written for users of statistical methodology, Ziliak and McCloskey have made it accessible to non-statisticians by citing published critiques of studies in a wide variety of disciplines that appear to throw viable yet statistically insignificant babies out with the bath water. As historians of economics, they are perhaps most persuasive when documenting the misuse of significance testing in various branches of economic science. Yet their studies of related disciplines, such as psychology, sociology, law, education, ecology, epidemiology, jurisprudence, and medicine, are almost equally persuasive.

All in all, Ziliak and McCloskey have written a fairly comprehensive history of the century-old dispute between those who follow Fisher in doubting the reality of any effect labeled "statistically insignificant," and those who follow Gosset in being prepared—under carefully delineated circumstances—to impute worldly significance to statistically insignificant effects. Tests of statistical significance are perforce unable, in Gosset's opinion, to deal with what he called "real error," which he often measured in units of "pecuniary advantage" or "clinical effect."

To illustrate the inadequacy of statistical significance testing, the authors reproduce part of an ad for Xanax, a product of the Upjohn Company designed to alleviate clinical anxiety. The ad, which ran in the May 1983 issue of the *Journal of Clinical Psychiatry*, contained a rather concise summary of the science behind Xanax, which compared the effects of the drug with those of diazepam (Valium), then the most widely prescribed anxiety drug.

Valium's most annoying side effect is drowsiness, especially among the depressed. The ad in question declared, among other things, that Xanax had a "significantly lower incidence of drowsiness when compared directly with diazepam therapy," and was "significantly" better than a placebo in decreasing the depressed mood scores among patients believed to be "significantly" depressed. Neither the ad nor the report to which it referred explained how much drowsiness diazepam patients experienced, or how much less of it—measured perhaps in hours of sleep reduction—afflicted those treated with Xanax. How big a difference is a significant difference? How much did the placebo decrease depressed mood scores, and how much more were they decreased by Xanax? How depressed is significantly depressed?

These, the authors insist, are precisely the sorts of questions one needs to ask when seeking to interpret the results of a clinical trial. MDs, they submit, should have access to such information (along with the prices of the two drugs) when deciding which, if either, to prescribe. Admittedly, Upjohn couldn't have shoehorned all that information into a single commercial message. But shouldn't underlying scientific reports answer such questions? Shouldn't "interested parties" be able to look up the responses somewhere? At present they rarely can, because detailed descriptions of clinical studies seldom pass into the open literature. Such information ordinarily stays behind (along with the "raw data") in investigators' files, which tend to be destroyed when the owners die, retire, or move to different institutions.

Similar remarks apply, as Gosset was quick to warn, to agricultural experiments comparing (say) alternative strains of barley. It isn't enough to know that strain A yields significantly more bushels per acre than strain B. One needs to know how much more A yields than B, in order to decide how long the new strain will take to repay the not inconsiderable investment involved in developing it. Responsible investment decisions cannot be based on statistical significance alone.

Perhaps the most frightening chapter in the book describes a 1978 study conducted by a team of doctors and statistical researchers at Mount Sinai Hospital in New York and published in the *New England Journal of Medicine*. The study analyzed 71 clinical trials that had been halted when it became clear that the benefits of the treatment being tested were "statistically insignificant." The study team concluded that 67 of the 71 trials had been terminated prematurely, when there was still a 10% chance that the treatments could in fact deliver therapeutic improvement of 25%; indeed, 50 of the trials were terminated when there was a 10% chance that the treatment could deliver a 50% therapeutic improvement.

Moreover, because a majority of the treatments tested were for cancer or cardiovascular or gastrointestinal ailments, the measure of therapeutic improvement was frequently the patient's chance of survival. In one of the trials, a treatment was abandoned while the probability of achieving a 25% reduction in mortality was still 77%, and the probability of a 50% reduction was still 42%! Such premature terminations—justified by statistical significance alone—were published in *The Lancet*, the *British Medical Journal*, the *New England Journal of Medicine*, the *Journal of the American Medical Association*, and other elite journals.

One might think that the existence of such studies—and Ziliak and McCloskey cite several, in a variety of disciplines—would launch a wave of methodological reform. Yet no such reform has been observed. Throughout the social and medical sciences, the river of statistical significance testing flows on unabated. Particularly in economics, the authors' home field, they found that dependence on statistical significance testing had actually increased from the 1980s to the 1990s.

If people read it and heed it, Ziliak and McCloskey could turn out to have written an exceedingly influential book. Otherwise, nothing will change. At present, the status quo seems to be holding firm. As with campaign finance reform, it is not enough that all should see and acknowledge the need for change. Where sacrifice is involved, there must also be incentives. To date, the few journals that have tried to revise their publication standards upward, so as to preclude mere tests of statistical significance, have not been widely imitated. Even accomplished practitioners of the statistical sciences are understandably reluctant to sacrifice the ease and familiarity of mechanical *t*-tests and *F*-tests for messier (and more laborious) measures of "oomph."

James Case writes from Baltimore, Maryland.