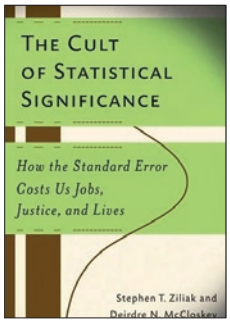
Praying to the power of *P*



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

Stephen T. Ziliak and Deirdre N. McCloskey

University of Michigan Press, 2008
352 pp., paperback, $24.95
ISBN: 0472050079

Reviewed by Jessica S Ancker

Imagine learning you have cancer. "But there's good news!" your doctor cries. "We have a new drug that has a *P* value of less than 0.01!"

Obviously, this statement is useless because it conveys no information about how well the drug works. What proportion of patients does it help? How much can you expect the tumor to shrink? What are the costs, in money or in side effects? That is the sort of information needed to decide whether the drug is worth taking, and it doesn't come from statistical testing. It comes from the data.

Unfortunately, many researchers assume that statistical testing will tell them whether their findings are important. Stephen T. Ziliak and Deirdre N. McCloskey have made it their mission to crusade against this assumption. In their new book, *The Cult of Statistical Significance*, they provide examples of inappropriate reliance on statistical testing, document bad statistical practice in a variety of scientific disciplines, review debates in the history of statistics and show repeatedly that 'statistical significance' has nothing to do with scientific significance.

To explain the main point of the book, it may be useful to review what statistical significance is and is not. Consider a study of a cancer drug in animals with tumors. If the question is whether the drug reduced tumor size by a meaningful amount, we can answer it only by looking at the average size decrease and deciding whether it is big enough to be meaningful. As Ziliak and McCloskey point out, statistical testing cannot answer this question, because the answer is a judgment that requires knowledge of the topic. For example, a tiny decrease might be important for a previously untreatable cancer, whereas complete remission might be the goal for a cancer that has other successful therapies.

Statistical testing is designed to answer a less interesting question: how likely is it that the experiment would produce these results purely by chance? Random fluctuations happen all the time by chance alone, and researchers might not want to be too quick to confuse this occurrence with a true drug effect. In fact, the mere fact that a sample was studied (instead of the entire hypothetical population in question) introduced random error, because samples differ from each other by chance—a phenomenon called sampling error. Performing a statistical test in our tumor study first requires estimating the sampling error with a measure called the standard error. We then compare the experimental results (the average tumor decrease) to the standard error. The *P* value is the probability of getting this particular ratio if we have a completely ineffective drug; that is, by pure chance. The *P* value will be small if the effect of the drug is much larger than the standard error, regardless of whether the effect has any importance. However, thanks to an arbitrary threshold set by statistics pioneer R.A. Fisher, the term 'significance' is typically reserved for *P* values smaller than 0.05. Ziliak and McCloskey, both economists, promote a cost-benefit approach instead, arguing that decision thresholds should be set by considering the consequences of wrong decisions. A finding with a large *P* value might be worth acting upon if the effect would be genuinely clinically important and if the consequences of failing to act could be serious.

Another excellent recommendation in the book is to avoid *P* values altogether and use confidence intervals instead. The confidence interval is a range of plausible extrapolations about a population from the sample data. The mean tumor decrease observed in the sample described above might be 5 mm, and a 95% confidence interval of, say, 2 mm to 8 mm provides a range of plausible values for what the mean decrease might be if the entire population could be studied instead of just the sample. The confidence interval, though transformable into a *P* value, has a very different psychological impact because it draws the reader's attention first to the data. It encourages the reader to ask such questions as how big was the decrease, how big were the tumors to begin with and how does this effect compare to the effect of other therapies? It also allows the reader to judge the precision of the estimate (is the interval wide or narrow?) and the implications of the interval (would I change my opinion if the result were somewhere else within this interval?). In other words, confidence intervals encourage meaningful qualitative judgments about quantitative data.

The book may not succeed in converting students or scientists unfamiliar with statistics, however, because it does not provide a basic review of the math of statistical testing and estimation, nor does it explain other methods mentioned in passing, such as Bayesian approaches. The book also seems likely to alienate many readers who do know something about statistics. For example, the authors blast a list of scientists who obviously understand the limitations of statistical testing simply because their articles contain *P* values. Such researchers are accused of pursuing "superstition" and "metaphysics" rather than science. The harsh tone, as well as the frequent italics, exclamation points, neologisms and nicknames (a disapproving "Wasp" for Fisher and an admiring "Bee" for his contemporary William Sealy Gosset) might be entertaining in an essay but become somewhat grating in a book 300-plus pages long. The appropriate use of statistics in science and policy is an important topic, and the authors make many good points, but their book might irritate more people than it persuades.

Jessica S. Ancker is in the Department of Biomedical Informatics, Columbia University, College of Physicians and Surgeons, 622 West 168th Street, VC5, New York, New York 10032, USA.
e-mail: jsa2002@columbia.edu