Ziliak, Stephen T. & Deirdre N. McCloskey – The Cult of Statistical Significance

The University of Michigan Press, 2008, [Surrounding Knowledge] Grade 

Suppose I told you I had found irrefutable scientific evidence that a CEO's golf handicap affects his or her company's stock performance. Better golfers imply higher stock returns, I'd say. As evidence I would produce a huge regression model, covering thousands of companies, where golf handicap as explanation of differing stock returns was found to be statistically significant at the 1 percent level. Now promise me you will never trust any such mock science.

Stephen Ziliak and Deirdre McCloskey, two economics professors with a keen sense of the right and wrong uses of statistics, have written a truly eye-opening book. Without resorting to equations, they show that the practice known as *hypothesis testing* or *tests of statistical significance* is utterly flawed. Unfortunately, most scientists are unaware of this fact and scientific journals of the highest standing continue using it anyway.

Let it be said at once, Ziliak and McCloskey's book is not an enjoyable read. Not only for its uncomfortable implications but just as much for its poor editing. On the good side the prose is sometimes beautiful and often funny. But its points are rarely well argued, the quotations and examples are too many and too lengthy and the endless repetition of its central polemic soon becomes tedious. However, I am willing to accept these flaws as the book is too important to be written off on the grounds of style.

The authors take a searing look at the practice of labelling factors either statistically significant or insignificant. The method is often associated with the statistician Ronald A. Fisher. But the authors show how Fisher virtually abducted the thinking from William Sealy Gosset, the head brewer of Guinness brewery, but corrupted Gosset's original ideas and presented them in a simplistic and flawed version. Gosset is the man behind the famous pseudonym Student which has given name to Student's t.

The biggest flaw in significance testing is to ask only *whether* there is a relation, not the *magnitude* of the assumed relation. Some factors could come out statistically significant while having negligible impact (like the golf handicap?), while others could have large impact while never achieving statistical significance. The outcome is to some extent in the hands of the researcher. Choose a big enough sample size and almost anything becomes statistically significant. Conversely, a small enough sample size makes nothing significant. The risk of manipulation by researchers is obvious as they can pick and choose between significance and insignificance by choice of sample size.

Hypothesis testing is even asking the wrong question. It asks what the likelihood would be of observing the data you have collected, assuming the null hypothesis is true. But the more relevant question concerns the likelihood of the hypothesis being true, given the observed data. This is equivalent to confusing the probability of a person being dead given that he was hanged, with the probability of him being hanged given that he is dead. Quite a difference, I would say.

The chilling implication of Ziliak's and McCloskey's book is that it casts doubt on almost everything we regard as knowledge in society. If papers published in eminent journals like the *American Economic Review* make such elementary errors, how are we to trust any scientific findings at all? Let alone any of the causes and effects communicated to us by less rigorous media?

Alf Riple, June 23 2013