STEPHEN T. ZILIAK AND DEIRDRE N. McCLOSKEY (2008), *The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor, University of Michigan Press, pp. 348.

The Neverending Story: A Note on Statistical Significance and its Critics

The debate over significance tests has surely been one of the longest-standing methodological controversies among statistically-minded social scientists. Indeed, Stephen T. Ziliak and Deirdre N. McCloskey's recent book-length assault on what these authors describe as a pervasive, faulty and dangerous practice, is only the last one of a considerable string of readers, articles, and symposiums devoted to this issue over the last half-century or so. The controversy has ranged across a number of disciplines, notably sociology, psychology, epidemiology and the health sciences in general, as well as biology/ecology. At a quarter of a century's distance, Morrison and Henkel's *The Significance Test Controversy* (1970) and Harlow, Mulaik and Steiger's *What If There Were No Significance Tests?* (1997) have offered collections that aptly and fairly documented respective positions and summarized the pros and cons of the issue without coming to any reconciliation. A bibliography compiled almost a decade ago (Thompson, 2001) enumerated no less than 402 titles devoted to the criticism of significance tests – yet it was not exhaustive (papers supporting their use were not included), and, according to a rapid survey this reviewer made, more than thirty-five other contributions at least have appeared since. In their own field of economics, Ziliak and McCloskey (from now on: Z&M), together or separately, have themselves been hitting hard on this nail for almost 25 years. In 2004, the *Journal of Socio-Economics* devoted a whole issue to a discussion of a Z&M paper entitled «Size Matters», and in which they presented the results of a survey of statistical practices of authors published in the *American Economic Review* (AER) over two decades. In *The Cult of Statistical Significance*, a title that echoes the deprecating tone often used by earlier critics who evoked «the sacredness of .05» and described the use of tests as a «ritual» that eliminated the need for thinking, Z&M vigorously pursue their campaign against what they call «the standard error» by coming back on their AER survey and rebuking their critics, documenting the outcomes of the debate in other fields (mainly psychology and medicine/epidemiology), and providing us with an historical sketch of how statistical significance became the gospel of statistical practice thanks to the evil doings of R. A. Fisher.

For a number of decades, significance testing, or, more precisely, null hypothesis significance testing, has indeed become a widespread and even compulsory procedure in a number of disciplines. The point of significance testing is to determine whether a result exhibited by a random sample still holds for the universe from which the sample is drawn: in other words, is the statistic or observed relationship genuine and susceptible of generalization or is it merely due to chance fluctuation? Significance testing is thus concerned with the management of error, in fact with that component of total error due to random variation. Basically, it consists in comparing what we observe with what we expect. Practically, it often consists in comparing what we observe with what is called the null hypothesis, generally understood as the hypothesis that there is no relationship between the variables. In fact, the researcher seeks to disprove the null hypothesis with a given level of confidence, from which she often concludes that her own hypothesis holds. In order to make up her mind about accepting or rejecting the null hypothesis according to some objective criterion, she has to determine a specific threshold: this threshold or significance level has traditionally been set at .05, which means that when an observed relationship would obtain in 95 out of a 100 similarly drawn random samples, she can reasonably hypothesize that it is not due to chance. Of course, she may still err because her sample may precisely be one of the 5 that may randomly exhibit such a result, but, having followed a mechanical procedure, she cannot be reproached with partiality in favour of her cherished views. Confidence will evidently grow if significance level reaches .01 or .001 (proudly accompanied by * and ** in statistical packages reports). As a result, it has become general practice, in most scientific journals, not to publish quantitative papers that do not submit their data to significance tests, and not to report data when the significance level has not reached the .05 threshold.

Now, according to Z&M, there are a number of problems here. The major one – at least with regard to economics – is that significance testing has focused attention on the existence of relationships and therefore diverted it from the importance or size of these relationships; in the author's words, it has substituted a philosophical question (whether?) to the crucial social-scientific concern of size or, as they say: *oomph* (how big?). This distorsion is embodied by the expression statistical *significance* itself, too easily confused with *substantive* significance, by which we may understand *economic* significance, *sociological* significance, *clinical* significance, etc. For a number of years, Z&M have criticized their fellow economists for precisely neglecting size effects, which they see as valid indicators of economic significance, and often satisfying themselves with reporting significance levels or the mere direction of relationships («sign econometrics») as well as committing a series of other sins. In a survey of «all full-length articles published in the *American Economic Review* (AER) in the 1980s (N = 182) and 1990s (N = 187)», they examined how these papers dealt with 19 relevant issues such as sample size, power of the test, the size of coefficients, the senses in which the word *significance* was used, the attention devoted to assessing the importance of observed relationships («how big is big?»), etc. (p. 81). The state of things they describe is quite troubling: for instance, no less than 72% of the papers published in the AER in the 1980s did not address this last question, 70% did not distinguish between statistical significance and economic or policy significance, and nearly 60% of them «used the word *significant* equivocally» (p. 75); the situation had not clearly changed in the 1990s, which saw improvement on some aspects and deterioration on others. Neglecting size effects and confusing statistical with substantive significance have unfortunate results: namely, they may focus attention on statistically significant yet small effects that yield no adequate understanding of the phenomenon under study, or lead to reject other, larger, and potentially important effects that do not meet the required .05 threshold due, for instance, to small sample size. As a dramatic instance of this, Z&M point to the infamous clinical trial of Vioxx, where the number of patients taking Vioxx that suffered from heart attack (5) was not statistically significant compared with that (only 1) of the control group (of course, besides the statistical issue, others, of an ethical character, such as having hidden three other cases of heart attack, were involved, that fall under the purview of law rather than of methodological critique). Another telling example of how importance (or size) is sacrificed to precision (or statistical significance) is provided by Ziliak's own experience as a labor market analyst for the State of Indiana. He reports that the fairly impressive unemployment rates of black youths (on average 30 to 40 percent but derived from small samples) were not made public because their *p*-values were too high (over .10). In other words, according to the gospel of statistical significance, no matter how impressive a finding may be, precision remains the criterion on which judgment should rest. What economists – and social scientists in general – lack, according to the authors, is a «loss function», a criterion against which to pass judgment that relates to the size of observed values (as in the case of black teenagers' unemployment rates) rather than to the relation between sample and universe. Now, that idea seems reasonably applicable in the domain of industrial production, where a decision has to be made regarding a course of action (for instance, stopping or maintaining the production line on the basis of compared expected losses in case of a right or a wrong decision) and it is therefore not surprising that J. Neyman, W. Shewhart, W. E. Deming and A. Wald are quoted, among others, in defence of it. Against the argument that Pure Science should not be conducted according to «cost», Z&M maintain that social science research has a fundamentally pragmatic character and should therefore always take into account what losses or gains might be entailed by a given course of action, something «the sizeless stare of statistical significance», concerned with existence rather than quantity, cannot provide. This view implies of course that decisions have to be made by the researchers themselves, in view of what others have done earlier, and that subjectivity should gain some ground at the expense of mechanical devices such as null hypothesis significance tests. It has for instance been proposed, in a paper not quoted in the book, that clinical, in the sense of substantive, significance may also be assessed through the statistical measurement of magnitude of effect (precisely Z&M's *oomph*) and, on a more qualitative basis, through social validation by asking patients about their sense of well-being (Lefort, 1993). Interestingly, these kinds of suggestions and the critique of significance testing in general have been shared by the more «politicized» quarters of the dismal science rather than by its mainstream, as shown by the editorial practice of *The Journal of Socio-Economics* or *Feminist Economics*, two journals that explicitly enjoin authors to clearly distinguish between statistical and analytical significance, and the favourable hearing Z&M have met with among their own libertarian family.

Z&M's original survey, whose results had first been published in the 2004 paper, has been fiercely attacked as badly flawed and sloppily conducted, and the claim that statistical significance is not economic significance has been described as «jejune and uncontroversial» (Hoover and Siegler, 2008: 1). But that survey merely tried to document for economics identical charges that have been made in other disciplines. Indeed, from the vantage point of 2008, Z&M are able to position their own contribution besides comparable, and sometimes much larger, surveys made by other critics of significance testing, mainly in epidemiology and psychology: the overall diagnosis is equally dire, even though attempts at reforming statistical practice have been partly successful in certain quarters. Whatever faults Z&M's survey of their fellow economists' practice may contain, it seems difficult to squarely reject their overall conclusions and pretend, as some of their critics have done, that everything is fine in the finest of all possible worlds. One of the remarkable aspects of the controversy is indeed the contrast between, on the one hand, its truly multi-disciplinary scope and, on the other, its disciplinary-contained character, which sees the same procedures criticized, the same arguments put forward, the same remedies proposed, the same crusades waged, not to much avail. (As a political scientist by trade, the present reviewer was curious of the echo the controversy had in his own discipline; it was in fact pretty dim, despite one essay that remarkably summarized all key aspects of the debate and, after having examined 148 articles [drawn from four major political science journals 1997 volumes] that used null hypothesis significance testing, found that 65 of them «drew substantive conclusions from a fail to reject decision» [Gill, 1999: 660-1]).

Now, besides criticising the actual statistical practice of economists, Z&M also draw their fire over the general logic of significance testing, restating, sometimes amusingly and vividly, arguments that had been made for other disciplines earlier, such as the illegitimacy of equating null hypothesis testing with the logic of *modus tollens* given the former's probabilistic nature, the fallacy of the transposed conditional, i.e. drawing from the inconsistency of an observation with the null hypothesis an argument in favour of our hypothesis (see the Gill 1999 quotation above), the bias of model selection, i.e. choosing between equally reasonable hypotheses on the basis of significance levels, or the problems related to sample size evoked above. (One fascinating issue they deliberately choose not to examine is the practice of conducting significance tests on data that does not originate from random sampling or on statistical universes envisioned as samples drawn from larger hypothetical universes.) Here again, the failure of these repeatedly asserted arguments to critically injure or permanently reform the practice of social scientists remains somewhat of a mystery, especially given the impressive string of authors Z&M present as holding the same views as theirs: is the persistence of significance testing due to habit and laziness, to the fact that a sub-optimal practice cannot easily be displaced once it has gained dominance (significance testing as a methodological analog of the QWERTY keyboard), or is the story more complicated?

Z&M's tentative answer to the question of how statistical significance testing came to dominate social scientific practice includes an interesting, even if sometimes questionable, excursus in the history of 20th-century statistics. The interesting aspect lays in the part of their narrative that explores how a technical innovation devised for very specific purposes (William S. Gosset's «probable error» and his *Student's t*-distribution) progressively evolved into a rigid practice that generated confusion between estimating the strength of a relationship and measuring the probability of its existence. In this story, a (substantively, of course) significant role is played by the politics of statistics as a discipline: R. A. Fisher's influence is described in the strategic idiom of empire-building, and the United States and India as conquered grounds thanks to the conversion of enterprising pupils such as Harold Hotelling and P. C. Mahalanobis. But more impersonal factors are at play, such as the appeal of High Modernism, which made Fisherian statistical testing benefit from «the prestige of the mechanical methods in all fields, from mathematics to automobile manufacturing» (243) and sets it in a rather large series of social-scientific schemes among which we may also include the scientific organization of labour, central economic planning (hence the interested attention of libertarians), or even – to follow James C. Scott, an author not quoted in the book but with whose views Z&M would have concurred – Le Corbusier's architecture. Closely linked to High Modernism is Robert K. Merton's concept of the Bureaucratization of Knowledge, which sees means gaining the status of ends while formalisms and rituals replace substantive inquiry (*ibid.*). The authors convincingly demonstrate that, from its beginnings and despite the dominance it gained, Fisherian significance testing has never gone uncontested in the field of statistical science itself and that their own efforts can boast of an impressive pedigree. What may sometimes irritate the reader is the moralistic tone that pervades the narrative. Not only has Fisher been the bad genius of statistics and taken us along a wrong and damageable path, but he also was, it seems, a thief and a counterfeiter (of Gosset's tables, which he appropriated and misinterpreted), and a very mean human being as well. The authors' account of the rise of statistical significance as «trained incapacity» therefore oscillates between, on the one hand, an explanation based on path dependency and the interplay of intellectual and sociological factors, and, on the other hand, the impulse to give a prominent role to fraud and deceit. These alternatives somewhat mirror those held on statistical significance testing by its critics and its defenders. According to the former, its generalized use in social science amounts to a totally misguided scientific practice, whose logical problems and ruinous consequences can be argued and documented. The latter respond that the only problem resides in lazy work on the part of unsophisticated practitioners or in outright fraud, as in the Vioxx case. Even though intellectual history eventually connects with individual behaviour, putting too much emphasis on an individual's character and what amounts finally to making a case upon supposed intentions somewhat weakens Ziliak and McCloskey's otherwise sound and elaborate critique. The protracted fighting over statistical significance goes on.

REFERENCES

Gill, Jeff. 1999. «The Insignificance of Null Hyothesis Significance Testing», *Political Research Quarterly*, Vol. 52, No. 3 (September), pp. 647-674.

Harlow, Lisa L., Stanley A. Mulaik and James H. Steiger, eds. 1997. *What If There Were No Significance Tests?* Mahwah (N.J.)/London, Lawrence Erlbaum Associates, pp. 446.

Hoover, Kevin D. and Mark V. Siegler. 2008. «Sound and fury: McCloskey and significance testing in economics», *Journal of Economic Methodology*, Vol. 15, No. 1 (March), pp. 1-37.

Lefort, Sandra M. 1993. «The Statistical versus Clinical Significance Debate», Journal of Nursing Scholarship, Vol. 25, No. 1 (Spring), pp. 57-62.

Morrison, Denton E. and Ramon Henkel, eds. 1970. *The Significance Test Controversy*, New Brunswick (N.J.)/London, Aldine Transaction, pp. 333.

Thompson, William L. 2001. «402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies». http://warnercnr. colostate. edu/~anderson/thompson1.html.

Ziliak, Stephen T. and Deirdre N. McCloskey. 2004. «Size matters: the standard error in the American Economic Review», *The Journal of Socio-Economics*, Vol. 33, pp. 527-546.

Jean-Guy Prévost