WP-2014-038

The Cult of Statistical Significance - A Review

Sripad Motiram

Indira Gandhi Institute of Development Research, Mumbai
September 2014
http://www.igidr.ac.in/pdf/publication/WP-2014-038.pdf

The Cult of Statistical Significance - A Review

Sripad Motiram
Indira Gandhi Institute of Development Research (IGIDR)
General Arun Kumar Vaidya Marg
Goregaon (E), Mumbai- 400065, INDIA
Email(corresponding author): sripad@igidr.ac.in

Abstract

I present a review and extended discussion of The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives by Deirdre McCloskey and Stephen Ziliak, a work that raises important issues related to the practice of statistics and that has been widely commented upon. For this review, I draw upon several other works on statistics and my personal experiences as a teacher of undergraduate econometrics.

Keywords: Significance; Standard Error; Application of Statistics; Methodology

JEL Code: C1; C12

Acknowledgements:

I thank Dilip Nachane for useful discussions and Om Narasimhan for providing me with some of the material. I have drawn upon the reactions of my students in an undergraduate econometrics course at Dalhousie University, and I also thank them.

The Cult of Statistical Significance – A Review[*]

Sripad Motiram[**]

Abstract

I present a review and extended discussion of *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives* by Deirdre McCloskey and Stephen Ziliak, a work that raises important issues related to the practice of statistics and that has been widely commented upon. For this review, I draw upon several other works on statistics and my personal experiences as a teacher of undergraduate econometrics.

JEL Codes: C1, C12
Keywords: Significance; Standard Error; Application of Statistics; Methodology

In 1710, British polymath John Arbuthnot[1] examined data on children in London for an 82-year period (1629-1710) to argue that not chance, but "divine providence" dictated sex-ratio at birth.[2] This is the earliest instance of what in modern parlance can be termed as a test of significance.[3] Apart from demography, significance testing was initially used in astronomy, e.g. in his 1773 memoir Laplace deployed it to investigate whether comets originated within our solar system (Scott 1953). In their fascinating recent book (*The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*) Stephen Ziliak and Deirdre McCloskey (ZM, hereafter) show how from these modest beginnings, significance testing has become central to many sciences today, and document the deleterious consequences of this phenomenon.

These arguments were made earlier by ZM and McCloskey in several papers, which led to an article in the *Economist* and a debate in the *Journal of Socio-Economics*. While in these papers, their main focus was on demonstrating the misuse of statistical significance in economics the book has a broader scope. It shows us how deep and pervasive this malaise is, affecting disciplines as diverse as animal science, economics, management and medicine, to name just a few. It provides a historical and intellectual context within which we can locate the rise and consolidation of the application of statistical techniques in general, and statistical significance in particular. It also provides guidance for practice in the future. All this is done in a

[*] I thank Dilip Nachane for useful discussions and Om Narasimhan for providing me with some of the material. I have drawn upon the reactions of my students in an undergraduate econometrics course at Dalhousie University, and I also thank them.

[**] Associate Professor, Indira Gandhi Institute of Development Research, Gen. A.K. Vaidya Marg, Goregaon (E), Mumbai, India-400065. E-Mail: sripad@igidr.ac.in.

[1] Arbuthnot was a physician and a satirist who collaborated with Jonathan Swift. He also published the first work on probability in English, wherein he translated and extended a tract of Christian Huygens, which itself was the first printed work on probability (Stigler 1986, p. 221).

[2] Arbuthnot (1710) argued that deaths among men are higher than the same among women since the former face hazards while seeking food. Hence, God restores balance by bringing forth more males than females, which if left to chance would almost surely not happen, given the low likelihood of this event. He thereafter concluded that polygamy is contrary to the law of nature!

[3] See Aldrich (2009), who also points out that the term "significant" was first used by Edgeworth in 1885, in discussing the difference between the mean stature of British criminals and the mean stature of adult males from the general population.

witty manner, using examples from popular culture, and even haikus. The book also clarifies (at least for me) some of the points that ZM had made in their earlier work.

The main message of the book is that there is a difference between statistical and substantive (i.e. policy/practical/economic/clinical etc.) significance. In most situations that one encounters in practice, and/or in disciplines such as economics or medicine, the relevant question to ask is whether an effect is "big" or "small." Statistical significance, which is really a measure of precision, does not address this question. Precision is a good thing to have, but is neither necessary nor sufficient for a finding to be important. ZM document many studies, in a variety of disciplines, which have missed this point, by confusing between these two notions of significance, by ignoring magnitudes, and by satisfying themselves with just sign and statistical significance. They document several other abuses of significance testing.[4] Using a 19-item questionnaire, they evaluate empirical papers published in the leading general interest journal in economics (*American Economic Review* (*AER*)) in the 1990's and 1980's and conclude that practice continues to be bad in economics, and has actually worsened in the 1990s.

I think that the main argument of the book is correct, and I will illustrate this with a simple example that I constructed to persuade myself. In a study of regional disparity in India, a hypothesis test of the equality of mean daily real wage rates in two neighboring states is conducted, at some arbitrary significance level (say 5 percent). Surely, in any situation where this study could be useful, we would be interested in knowing how large the difference between the wage rates is, i.e. whether it is "big" or "small." Merely knowing whether the difference is statistically significant or not, does not help us. Is a difference of Re 1, although statistically significant, practically important? It depends. If the Indian government wants to reduce disparity, it may not be worth the trouble, i.e. the difference may be too small to worry about. On the contrary, for a poor laborer near the border, a Re 1 difference, whether it is statistically significant or not, could be quite important. Incidentally, as the formula for standard error tells us, with large samples, even minor differences from a practical point of view can become statistically significant.

This message is not new - as ZM themselves acknowledge, it was known to several distinguished economists (e.g. Keynes, Edgeworth, Arrow), statisticians (e.g. Gosset, Neyman, Kruskal) and other scientists (e.g. Deming). Moreover, many of the practices that ZM identify are clearly abuses (e.g. reporting all test statistics, whether relevant or not) and some (e.g. selecting variables solely based upon statistical significance) have even been explicitly advised against in undergraduate textbooks (Gujarati 2003, Ch. 13). So, overall, what ZM say is uncontroversial. Therefore, in the debate[5] that arose on this issue (summarized in the book), which involved some leading statisticians and econometricians, there was agreement on the substantive points. The only harsh criticism came from Hoover and Siegler (2008). Even they agree with ZM's main message, but dispute the argument that there is misuse of statistical significance in economics. My reading of the exchange between them and

[4] e.g. Reporting all test statistics, whether they are necessary or not; ranking variables on the basis of their test statistics; choosing variables for inclusion solely based upon statistical significance etc.

[5] Most articles were published in the *Journal of Socio-Economics*, but there was also some discussion in the *Journal of Economic Methodology*, and the *Economic Journal Watch*.

ZM, and my own (limited) experience of a decade or so in economics, persuade me that ZM make a convincing argument.

One problem that ZM talk about at length is worth mentioning here. They point out that many studies using significance tests do not discuss the second kind of (Type II) error, and consequently use tests with low power. They provide an extensive discussion of this in the context of psychology, and also show that economists rarely talk about power.[6] Is this a serious issue? My own thinking is that it is. The approach to testing that underlies most studies is: posit null and alternative hypotheses; assume that a Type I error is more serious than a Type II error, and therefore choose a small (arbitrary) significance level; choose the best (most powerful) test.[7] Power can then be analyzed by looking at the power function - the probability of not committing a Type II error as a function of various assumed-to-be-true hypotheses (see e.g. Mood et al. 1974, Chapter IX). One can argue against this paradigm, but if one uses it, discussing power is important, since there is a trade-off between Type I and Type II errors. Also, the practical consequences of ignoring power can be quite telling. For example, suppose that a cancer drug is prevented from entering the market because a hypothesis that the drug is ineffective (i.e. as effective as a placebo) is tested, and not rejected. If the power of the test is low, what this means is that with a high probability a drug that could have saved lives is disallowed. ZM argue that economists rarely discuss power because in general, they do not understand it. I don't know whether this is true, but although I taught power functions to undergraduate students, till recently, I did not fully appreciate the importance of discussing power in practical contexts.

ZM argue, using several examples, that this is not just an academic issue (of research, academic careers and publications) but more serious – in some cases, literally a matter of life and death. An example that they provide (pp. 28-31) is quite illustrative. Vioxx, an arthritis drug was withdrawn by Merck (its manufacturer) after the death of a user, which raised concerns about its contribution to the risk of a heart attack. In the clinical trial, there were five cases of heart attack among Vioxx users, as compared to one among users of the generic drug, but Vioxx was still allowed because the difference (5 to 1) was not statistically significant at 5%. To date, several instances of heart attack among users of Vioxx have occurred, and Merck is straddled with lawsuits. The difference between 5 and 1 was not statistically significant, but it seems that it was clinically significant. Instead of mechanically applying the 5% rule, a proper consideration of the trade-offs by Merck and/or the regulatory authority (FDA) could have probably prevented this.[8] A second example (pp. 31-32) has a hint of dark humor – the Japanese government wanted to increase the number of whales that are hunted in Antarctica because this would give them larger samples (for research using whales) and thereby statistically significant results. I came across another example, viz., a vaccine against AIDS. The controversy here is whether we should accept results from a study where the effect of the vaccine is statistically insignificant.

---

[6] Only 8% (4.4%) of the *AER* papers that ZM looked at, in the 1990's (1980's), discussed power.

[7] This is a hybrid of the Fisher and Neyman-Pearson approaches. Fisher denied the necessity for an alternative hypothesis, or a loss function. On the contrary, Neyman and Pearson were very clear that the choice of the significance level should be based upon the trade-offs involved in a particular context (and not arbitrary). For a discussion of these issues, see Spanos (1999, p. 691).

[8] ZM also note that there are allegations of fraud - three patients using Vioxx, who suffered from heart attacks were dropped from the sample. If they had been considered, the difference would have been statistically significant.

Advocates of the vaccine are arguing that these results show the vaccine's effectiveness, even if there is lack of statistical significance, whereas opponents are arguing otherwise (Berkley 2009).

How did we get into this mess? The authors answer this question by tracing the history of statistics since the late nineteenth century, using both secondary and archival sources. I found this to be the most fascinating part of the book. The key figures here are Galton, Karl Pearson, Gosset and Fisher, and to a lesser extent Edgeworth, Egon Pearson, Neyman and Hotelling. The rise of the application of statistics (especially statistical significance) is located within the larger context of certain intellectual and philosophical currents: positivism, faith in science, neo-Darwinism, trust in numbers,[9] and even racist anthropometry. The main protagonist of this story is William Sealy Gosset, a brewer at Guinness, who discovered the t-distribution (apart from other things, e.g. power, crucial aspects of the design of experiments), who always emphasized substantive significance, and the ability of statistical tools to contribute to his main objective – brewing the best tasting beer. The culprit is Ronald Aylmer Fisher, who despite being a great scientist put statistics on the wrong track. Fisher promoted a mechanical 5% philosophy and denied the necessity for a loss function, or the second kind of error. Once set in motion, this message was propagated around the world by his disciples (e.g. Hotelling) and reproduced in countless number of texts. The personalities of Gosset and Fisher are also contrasted. While the former was humble, good-natured and a decent human being, the latter was a jealous and arrogant person, who tried to use his influence to stifle the careers of people who disagreed with him (e.g. Neyman), and who tried to take credit for the work of others (e.g. Gosset). I am not an expert on the history of statistics, and have not seen the primary sources that ZM (and others) have drawn upon. But, being somewhat of a keen student of the history of mathematics, I have read several works on the history of statistics (e.g. Stigler 1986, 1999), biographical sketches of Fisher, Gosset and other statisticians (e.g. Mahalanobis 1964, Spanos 1999, Rao and Székely 2000, JASA 2007). This part of the book complements these works. Fisher has many admirers in statistics and biology and they present a different picture of him (compared to ZM). Relatively speaking, Gosset's contributions seem to be underappreciated. Maybe, ZM restore the balance.

In economics, ZM argue that the kind of mathematical formalism pioneered by Samuelson, which emphasizes existence/qualitative thinking, abetted the rise of statistical significance (apart from other sins). ZM also voice their dissatisfaction with another related aspect of the economics profession, viz. the lack of pluralism (p. 240). Their main influence in advocating pluralism as a desirable world-view and as a methodological position (in statistical analysis too) is the work of philosopher Paul Feyerabend, whose influence can also be seen in other works of McCloskey. This is not the place to go into the merits and demerits of Feyerabend's controversial position (Broad 1979), which is an application of the principles of anarchism to philosophy of science (Feyerabend 1984, Ch 1). Notwithstanding criticisms of Feyerabend, I share ZM's dissatisfaction - unlike other social sciences (e.g. history, anthropology) in economics, there is one dominant (neoclassical) perspective and hardly any debate

[9] Maybe best expressed by Lord Kelvin: "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind" (Stigler 1986, p. 1).

among various perspectives. Although, I am friendlier to mathematical formalism than ZM, I see their point. Some other eminent economists (e.g. Coase, North) have been making similar critiques. My own view is that the economics profession can make better use of mathematics (as a tool, or language) than it is doing currently, by following the sane advice of one of its founders, viz. Marshall.[10] I think (hope) that the recent financial crisis has seriously exposed both these limitations. On the inadequacy of econometrics texts, that ZM refer to, I think there has been improvement. I found some recent texts (there may be others) that have discussed the issues that ZM raise.[11]

Where to go from here? In their final chapter, ZM discuss what is to be done to improve statistical practice. Their main advice is that people discuss substantive, practical significance of their findings and not commit the "error of undue attention," i.e. trying to solve a scientific problem using statistical significance and insignificance only. In terms of institutional changes, they suggest that journal editors impose a "Statement on the Proprieties of Substantive Significance," which requires talking about substantive significance, avoiding an arbitrary minimum significance level,[12] reporting sampling variance in some form (e.g. using confidence intervals), discussing power and the trade-offs involved in the hypothesis test (e.g. using Wald's loss function).[13] If implemented, I think that these suggestions will substantially improve quality of research and policy. The 19-item questionnaire that ZM used to score *AER* papers is itself I think a good guide to practice, although I would add that it should not be used mechanically.

Overall, ZM have written an interesting book that should be useful to anyone who teaches or uses statistics, or is plainly interested in mathematical statistics. I have used ZM's papers in an undergraduate econometrics class: for some students it was just a welcome diversion from dry material, but many others got the point and appreciated it. ZM's papers were useful for improving my own research, although I now realize that there is scope for further improvement. There is a clear Bayesian bias in some portions of the book, which might not be appealing to some people. I have not used the Bayesian approach in my own research (at least not as yet), but do find it insightful. However, I can also see that reasonable people can disagree with it. While everyone agrees that in mathematical terms probability is simply a function that

[10] Marshall had written in a letter to Bowley: "But I know I had a growing feeling in the later years of my work at the subject that a good mathematical theorem dealing with economic hypothesis was very unlikely to be good economics: and I went more and more on the rules - (1) use mathematics as a short hand language, rather than as an engine of enquiry. (2) Keep to them till you have done. (3) Translate into English (4) Then illustrate by examples that are important in real life. (5) Burn the mathematics. (6) If you can't succeed in four, burn three. This last I did often…I think you should do all you can to prevent people from using mathematics in cases in which the English language is as short as the mathematical." (Weintraub 2002, p. 22).

[11] e.g. Gujarati's popular undergraduate text *Basic Econometrics* refers to the difference between economic and policy significance (citing ZM) and also talks about the trade-offs involved in choosing a significance level. In their text, *Introduction to Econometrics*, James Stock and Mark Watson argue that the choice of the significance level should depend upon the context.

[12] This has also been advocated by others. The reporting of p-values is driven (at least partly) by the idea that different users could have different tolerance for Type-I errors.

[13] What is relevant here is that in most cases where significance testing is used, a decision is being taken, e.g. should a drug be introduced or not? Hence, thinking about trade-offs becomes necessary, even if one agrees with the claim (e.g. Spanos 1999, p. 570) that hypothesis testing can be inferential, rather than decision-theoretic in nature.

satisfies certain (Kolmogorov's) axioms, the philosophical interpretation of probability is contested terrain. The Bayesian subjective/degree of belief view is one among several (e.g. logical, frequency) views (Hajek 2009). There are other differences of methodology between Bayesian and rival perspectives (Lancaster 2004, pp. 1-4; Spanos 1997, pp. 568-70) on which there can be disagreement. But, this does not distract from the main point of the book, which is correct, and worth reiterating.

References

Aldrich, John. 2009. *Earliest Known Uses of Some of the Words of Mathematics (S),* http://jeff560.tripod.com/s.html

Arbuthnot, John. 1710. "An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes," reprinted in M.G. Kendall and R. L. Plackett (eds.), *Studies in the History of Statistics and Probability Volume II*, London: Griffin, 1977, pp. 30–34.

Berkley, Seth. 2009. "Have faith in an AIDS Vaccine." *New York Times*, October 18.

Broad, William. 1979. "Science and the Anarchist," *Science*, 206 (4418), pp. 534-537.

Economist (2004). "Signifying Nothing: Too many economists misuse statistics," Jan 29, 2005.

Feyerabend, Paul. 1984. *Against Method (3rd ed.)*, London: Verso.

Gujarati, Damodar. 2003. *Basic Econometrics (4th ed.)*, New York: McGraw Hill.

Hájek, Alan, 2009. "Interpretations of Probability," *The Stanford Encyclopedia of Philosophy* (ed.) Edward N. Zalta, http://plato.stanford.edu/archives/spr2009/entries/probability-interpret/

Hoover, Kevin and Mark Siegler. 2008. "Sound and Fury: McCloskey and Significance Testing in Economics," *Journal of Economic Methodology*, 15(1), pp. 1-37.

Journal of the American Statistical Association (JASA) 2008. "Discussion on Student's 1908 Article – The Probable Error of a Mean," 103(481), pp. 1-20.

Journal of Socio-Economics. 2004. "Discussion on Statistical Significance," 33(5).

Lancaster, Tony. 2004. *Introduction to Modern Bayesian Econometrics.* New York: Wiley-Blackwell.

Mahalanobis, P.C. 1964. "Some personal memories of R.A. Fisher," *Biometrics*, 20(2), pp. 368-371.

Mood, Alexander M., Franklin A. Graybill and Duane C. Boes. 1974. *Introduction to the Theory of Statistics*, New York: McGraw Hill.

Rao, C.R. and Gábor J. Székely. 2000. (ed.) *Statistics for the 21st century: methodologies for applications of the future*, New York: Marcel Dekker.

Scott, Elizabeth. 1977. "Testing Hypotheses," in Robert J. Trumpler and Harold F. Weaver (ed.) *Statistical Astronomy*, New York: Dover, pp. 220-230.

Spanos, Aris. 1999. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge: Cambridge University Press.

Stigler, Stephen. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge: Harvard University Press.

Stigler, Stephen. 1999. *Statistics on the Table*, Cambridge: Harvard University Press.

Weintraub, Roy. 2002. *How Economics Became a Mathematical Science*, Durham: Duke University Press.