*Erasmus Journal for Philosophy and Economics,*
*Volume 1, Issue 1,*
*Autumn 2008, pp. 154-164.*
*http://ejpe.org/pdf/1-1-br-2.pdf*

# Review of Stephen T. Ziliak and Deirdre N. McCloskey's *The cult of statistical significance: how the standard error costs us jobs, justice, and lives.* Ann Arbor (MI): The University of Michigan Press, 2008, xxiii+322 pp.

ARIS SPANOS*
*Virginia Tech*

The stated objective of this book is to bring out the widespread abuse of significance testing in economics with a view to motivate the proposed solution to the long-standing problem of *statistical vs. substantive significance* based on re-introducing 'costs and benefits' into statistical testing. The authors strongly recommend returning to the decision-theoretic approach to inference based on a 'loss function' with Bayesian underpinnings, intending to ascertain substantive significance in terms of "oomph, a measure of possible or expected loss or gain" (Ziliak and McCloskey 2008, 43).

The idea of a 'loss function' was introduced by Wald (1939), but rejected later by Fisher (1955) who argued that when one is interested in the truth/falsity of a scientific hypothesis, the cost of any actions associated with the inference is irrelevant; this does not deny that such costs might be relevant for other purposes, including establishing a range of substantive discrepancies of interest. This is still the prevailing view in frequentist statistics, which, to use one of the authors' examples (Ziliak and McCloskey 2008, 48), rejects the argument that to evaluate the substantive discrepancy from the Newtonian prediction concerning the deflection of light by the sun, one needs a loss function which reflects the relevant 'costs and benefits'.

How do the authors justify wedging the notion of a loss function back into econometrics? They interpret it in terms of 'economic cost' and trace the idea back to Gosset (1904); described pointedly as "a lifelong Bayesian" (pp. 152, 158, 300). How do they make their case? Curiously enough, not by demonstrating the effectiveness of their recommended procedure in addressing the statistical vs. substantive significance problem using particular examples where other 'solutions'

* **AUTHOR'S NOTE:** I am grateful to Kevin D. Hoover for many valuable comments and suggestions.

have failed. Indeed, in 320 pages of discussion, there is not a single credible illustration of how one can apply their proposed 'solution' to this problem. Instead, they attempt to make their case using a variety of well-known *rhetorical strategies and devices*, including *themes* like battles between good vs. evil, and conceit vs. humility, frequent *repetition* of words and phrases like 'oomph', 'testimation', 'sizeless stare' and 'size matters', picturesque language, metaphor and symbolism, flashback, allusion, parody, sarcasm, and irony. Their discourse in persuasion also includes some 'novel' devices like cannibalizing quotations by inserting their own 'explanatory' comments to accommodate their preferred interpretation, 'shaming' notable academics who 'should have known better', and recalling private conversations as well as public events where notable adversaries demonstrated the depth of their ignorance.

Their main plot revolves around a narrative with several ostensibly corroborating dimensions:

A. Evidence for the chronic abuse of statistical significance in economics.
B. Tracing the problem in statistics and the social sciences.
C. A 'selective' history of modern statistical thought as it pertains to the problem.
D. Discussion of various philosophical/methodological issues pertaining to the problem.
E. A 'what to do' list of recommendations to address the problem.

I will comment briefly on A-C and then focus my discussion on the last two dimensions.

**A**. The authors' accumulated evidence (chapters 6-7) for the widespread confusion between statistical and substantive significance in the abuse of significance testing takes the form of updating their 1996 scrutiny of applied papers published in the *American Economic Review* in the 1980s, which was based on grading these papers on 19 questions they devised for diagnosing the various facets of the problem. Although most of these questions are highly problematic in themselves, for the purposes of this review I will (reluctantly) take their evidence at face value and assume that most researchers sidestep the problem because they are unaware of a credible way to address it. Indeed, the researchers who scored very high on the M-Z scale only demonstrated *awareness* of the problem, but none of them, as far as I can see, had a credible

procedure to ascertain the substantive significance warranted by the data in question.

**B**. The literature on the problem of statistical vs. substantive significance is almost as old as modern statistics itself, and the authors do make an effort to trace its history all the way back to Edgeworth (1885) by stretching the truth somewhat to fit their narrative (see Hoover and Siegler 2008). Since the dominating objective for the authors is persuasion, this historical retracing is spread into several chapters (4, 10, 11, and 12) for impact, and as a result, it becomes rather diffused and less informative. The gist of the discussion is that, despite its long history, this problem has been raised in economics rather belatedly, and the authors do deserve some of the credit for making an issue of it, even though their discussion obfuscates the issues involved.

**C**. The narrative concerning the historical development of modern frequentist statistics which 'accommodates' their preferred interpretation of the problem is summarized as follows:

> We want to persuade you of one claim: that William Sealy Gosset (1876–1937)—aka "Student" of Student's t-test—was right and that his difficult friend, Ronald A. Fisher was wrong. […] Gosset, we claim, was a great scientist. He took an economic approach to the logic of uncertainty. For over two decades he quietly tried to educate Fisher. But Fisher, our flawed villain, erased from Gosset's inventions the consciously economic element. We want to bring it back (Ziliak and McCloskey 2008, xv).

Throughout this book, Fisher is painted as the villain of the story and Gosset as the patron saint of modern statistics whose contributions have been overlooked as a result of concerted efforts by Fisher and his disciples. Gosset (an employee of the Guinness brewery) is presented as the source of numerous great ideas in statistics which Fisher (a famed professor) was systematically embezzling while peeling off their 'economic element' (Ziliak and McCloskey 2008, xv). One such idea, as their story goes, was the evaluation of inferences in terms of their 'economic costs', and not the relevant error probabilities as such. Unfortunately for science, Fisher's conception of statistics prevailed, and Gosset's vision was forgotten by both statisticians and economists. One of the book's main objectives is to redress that.

It does not take much effort to discredit their narrative concerning Fisher and his role in the development of modern statistics because its inaccuracies and distortions are legion. The narrative reads like a

regurgitated but disconnected fable with Bayesian undertones; its heroes are primarily Bayesian 'at heart' and its villains are mainly Fisherian in perspective. However, even a glance through Savage (1976), one of the heroes, undermines the credibility of their narrative:

> *Just what did Fisher do in statistics?* It will be more economical to list the few statistical topics in which he displayed no interest than those in which he did. […] Fisher is the undisputed creator […] of the modern field that statisticians call the design of experiments, both in the broad sense of keeping statistical considerations in mind in planning of experiments and in the narrow sense of exploiting combinatorial patterns in the layout of experiments (Savage 1976, 449-450).

Acknowledging Fisher's epoch-making contributions to modern statistics does not, in any way, devalue Gosset's pioneering role in founding the frequentist approach in finite sampling theory, and influencing the work of both Fisher and Egon Pearson with insightful ideas and questions (see Plackett and Barnard 1990).

To illustrate the inaccuracy of the authors' narrative, let me simply oppugn one overhasty claim, that Arthur Bowley was a messianic disciple of Fisher who contributed significantly to spreading his statistical 'gospel' to economics (Ziliak and McCloskey 2008, 235, 293). Fisher revolutionized statistical thinking in the early 1920s while he was a non-academic statistician at Rothamsted Experimental Station; his first academic job, as professor of 'eugenics' at University College (London), was in 1933. Indeed, the academic establishment, led by Bowley (second only to Karl Pearson in academic status), fought with ferocity against Fisher's ideas, averted his appointment to several academic positions, and precluded him from most statistical forums, including the Royal Statistical Society (RSS). When this establishment could no longer ignore Fisher, Bowley and his cronies invited him to address the RSS for the first time in 1934, but their real intention was to expose him as a charlatan (see the discussion in Fisher 1935; and Box 1978).

**D-E**. The formal apparatus of the Fisher-Neyman-Pearson approach to frequentist inference was largely in place by the late 1930s, but its philosophical foundations left a lot to be desired. Several foundational problems, including: (a) the fallacies of acceptance and rejection, (b) the notion of statistical adequacy, (c) the role of substantive information in statistical modeling, and (d) the role of pre-data vs. post-data error probabilities (Hacking 1965), were left largely unanswered (Mayo 1996;

Spanos 1999). In particular, neither Fisher's p-value, nor Neyman-Pearson's 'accept/reject' rules, provided a satisfactory answer the basic question: 'When do data $\mathbf{x}_0$ provide evidence for or against a (substantive) hypothesis or claim?'

Indeed, both approaches are highly susceptible to:

(I). *the fallacy of acceptance*: (mis)-interpreting accept $H_0$ [*no* evidence *against* $H_0$] as evidence *for* $H_0$,
(II). *the fallacy of rejection*: (mis)-interpreting reject $H_0$ [evidence *against* $H_0$] as evidence *for* $H_1$; the best example of this is conflating statistical with substantive significance.

This created a lot of confusion in the minds of practitioners concerning the appropriate use and interpretation of frequentist methods. In the absence of any guidance from the statistics literature, practitioners in different applied fields invented their own favored ways to deal with these issues which often amounted to misusing and/or misinterpreting the original frequentist procedures (see Gigerenzer 2004). Such misuses/misinterpretations include, not only the well-known ones relating to the p-value, but also: (i) the observed confidence interval, (ii) the p-value curves, (iii) the effect sizes, (iv) the fallacy of the transposed conditional, (v) Rossi's real type I error, (vi) Zellner's random prior odds, and (vii) Leamer's extreme bounds analysis.

It can be argued that the authors' high-pitched recommendation of (i)-(vii), in their 'what to do' list to address the problem of statistical vs. substantive significance (Ziliak and McCloskey 2008, chapter 24), constitutes a perpetuation of the same foundational confusions, colored by the authors' Bayesian leanings, which have bedeviled frequentist inference since the 1950s. Space limitations prevent me from repudiating (i)-(vii) in any detail. Very briefly, the primary confusion underlying (i)-(ii) stems from the fact that, although observed confidence intervals do "draw attention to the magnitudes" (p. 73), they are no more informative on substantive significance than p-values; actually, there is a one-to-one mapping between the two, and they are equally vulnerable to the 'large $n$ [sample size] problem'. Moreover, the relevant post-data error probabilities in estimation are either zero or one—the observed confidence interval either includes or excludes the true value of the unknown parameter $\theta$—because the underlying reasoning is *factual* (under the true state of nature), as opposed to *hypothetical* (under different hypothetical scenarios) in testing.

The lack of proper post-data error probabilities in estimation explains why the different values of $\theta$ within an observed confidence interval are treated on a par, and the various 'effects sizes' proposed in the literature cannot possibly provide a reliable measure of substantive significance. Hence, the use of p-value curves to discriminate among the different values of $\theta$ within an observed confidence interval, giving the impression of attaching probabilities to these values (Ziliak and McCloskey 2008, 185), represents a mix-up of two different types of reasoning resulting in obfuscation (Spanos 2004). The charge that error probabilistic reasoning suffers from the fallacy of the transposed conditional stems from a false premise that error probabilities are conditional; there is *nothing* conditional about the evaluation of tail areas under different hypothetical scenarios, unless one conflates that with Bayesian reasoning which is conditional (Spanos 1999).

Among the various 'unsuccessful' attempts to address the problem of statistical vs. substantive significance that the authors dismiss, as yet another 'sizeless stare', is Mayo's (1996) post-data *severity evaluation* of the Neyman-Pearson 'accept/reject' decisions:

> If one returns to Mayo's discussion of what constitutes a "severe test" of an experiment, one finds only sizeless propositions, with loss or error expressed in no currency beyond a scale-free probability. […] A notion of a severe test without a notion of a loss function is a diversion from the main job of science, and the cause, we have shown, of error" (Ziliak and McCloskey 2008, 147).

It is clear from this quotation that the authors did not understand the use of this post-data evaluation in addressing the problem. *First*, contrary to their charge, there is no such thing as a 'severe test of an experiment', there are only severe tests of hypotheses or claims based on a particular test $T_\alpha$ and data $\mathbf{x}_0=(x_1,\ldots,x_n)$. *Second*, the severity evaluation, far from being another 'sizeless proposition', is actually framed in terms of a *discrepancy* parameter $\gamma \geq 0$ from the null, say: $H_0$: $\theta=\theta_0$ vs. $H_1$: $\theta > \theta_0$. The relevant post-data error probabilities—which remain firmly attached to the inference procedure itself and not to the hypotheses—evaluate the extent to which a substantive claim, such as $\theta \leq \theta_0 + \gamma$ or $\theta > \theta_0 + \gamma$ (associated with accept or reject), is warranted on the basis of a particular test $T_\alpha$ and data $\mathbf{x}_0$.

Depending on whether the Neyman-Pearson test has accepted (rejected) $H_0$, the severity evaluation is framed in terms of the smallest

(largest) warranted discrepancy $\gamma \geq 0$, measured on the same scale as $\theta$, with its magnitude easily assessable on *substantive* grounds. Hence, contrary to the authors' charge, the post-data severity evaluation of an accept/reject decision, gives rise to warranted discrepancies $\gamma$, which, in conjunction with substantive information, can help to address the fallacies of acceptance/rejection (see Mayo and Spanos 2006). Let me illustrate this.

**Example 1**. Consider the case where data $\mathbf{x}_0$ constitute a realization from the simple Normal model where $X_k \sim NIID(\mu,\sigma^2)$, $k=1,2,\ldots,n$. The t-test based on $\tau(\mathbf{X}) = \sqrt{n}\left(\overline{X}_n - \mu_0\right)/s$ is a UMP test for the hypotheses: $H_0: \mu = \mu_0$ *vs.* $H_1: \mu > \mu_0$ (see Cox and Hinkley 1974).

Assuming that $\overline{x}_n = .02,\ n = 10000,\ s = 1.1,$ yields $\tau(\mathbf{x}_0) = 1.82$, which leads to rejecting the null $\mu_0=0$ at significance level $\alpha=.05$, since $c_\alpha=1.645$. Does this provide evidence for a substantive discrepancy from the null? The post-data evaluation of the relevant claim $\mu > \gamma$ for different discrepancies $\gamma \geq 0$, based on $SEV(\tau(\mathbf{x}_0); \mu > \gamma) = P(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \gamma)$ [table 1], indicates that for a high enough severity threshold, say .9, the maximum warranted discrepancy is $\gamma < .006$.

| **Table 1: Severity evaluation of the claim:** $\mu > \gamma$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma$ | .001 | .005 | .006 | .01 | .02 | .05 | .07 |
| **SEV**$(\tau(\mathbf{x}_0); \mu > \gamma)$ | .958 | .914 | .898 | .818 | .500 | .003 | .000 |
| **POW**$(\tau(\mathbf{X}); c_\alpha; \mu = \gamma)$ | .060 | .117 | .136 | .231 | .569 | .998 | 1.00 |

One then needs to consider this in light of substantive information to assess whether the warranted discrepancy $\gamma < .006$ is substantively significant or not. In addition, the severity reasoning can be used to elucidate certain *fallacious claims* repeated by the authors throughout this book, pertaining to the very problem that occupies center stage: "A good and sensible rejection of the null is, among other things, a rejection *with high power*" (Ziliak and McCloskey 2008, 133). And "refutations of the null are easy to achieve if power is low or the sample is large enough" (p. 152).

No! No! You have it backwards. Rejection with high power is actually the main source of the problem of statistical vs. substantive significance, and 'large enough sample sizes' $n$ go hand in hand with high power, not low. For instance, the power of the above t-test increases with the non-centrality parameter $\delta = \sqrt{n}(\gamma)/\sigma$, which is a

monotonically increasing function of $n$. When a test has very high power for tiny discrepancies from the null, as in the large $n$ case, rejection of the null provides less (not more) evidence for the presence of a substantive discrepancy. This is illustrated in table 1, where the power of the test, based on POW($\tau(\mathbf{X})$; $c_\alpha$; $\mu = \gamma$) = P($\tau(\mathbf{X}) > c_\alpha$; $\mu = \gamma$), is very high for small discrepancies from the null; it is almost 1 at $\gamma$ =.05. What is even more misleading is that the power increases with the discrepancy $\gamma \geq 0$, in contrast to the severity evaluation.

Analogously, when a test with very low power for sizeable discrepancies of interest rejects the null, it provides more (not less) evidence for the presence of a substantive discrepancy.

**Example 2**. Let us consider the case where data $\mathbf{x}_0$ in example 1 yielded instead $\bar{x}_n$ = .633, $s$=1.1, for $n$=10; small sample case. In this case $\tau(\mathbf{x}_0)$= 1.82 leads to *accepting* the null $\mu_0$=0 at $\alpha$=.05 since the critical value now is $c_\alpha$=1.833. Does this provide evidence for *no* substantive discrepancy from the null? The post-data evaluation of the relevant claim $\mu \leq \gamma$ for different discrepancies $\gamma$, based on SEV($\tau(\mathbf{x}_0)$; $\mu \leq \gamma$) = $\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu > \gamma)$, indicates that for a high enough threshold, say .9, the minimum discrepancy warranted by data $\mathbf{x}_0$ is $\gamma$ >1.1.

| **Table 2: Severity evaluation of the claim:** $\mu \leq \gamma$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma$ | .1 | .25 | .5 | 1.0 | 1.1 | 1.2 | 1.5 |
| **SEV**($\tau(\mathbf{x}_0)$; $\mu \leq \gamma$) | .080 | .150 | .356 | .841 | .894 | .931 | .983 |
| **POW**($\tau(\mathbf{X})$; $c_\alpha$; $\mu = \gamma$) | .078 | .147 | .351 | .838 | .892 | .930 | .982 |

Again, substantive information should be used to assess if such a discrepancy is substantively significant or not. These two examples demonstrate how the same test result $\tau(\mathbf{x}_0)$=1.82, arising from two different sample sizes, $n$=10000 and $n$=10, can give rise to widely different 'severely passed' claims concerning the warranted substantive discrepancy: $\gamma$ < .006 and $\gamma$ > 1.1, respectively. Note that in the case of 'accept $H_0$' shown in table 2, the power moves in the same direction as severity and the two are close because $\tau(\mathbf{x}_0)$=1.82 is very near the critical value $c_\alpha$=1.833.

**Statistical adequacy**. Another inveterate foundational problem associated with the Fisher-Neyman-Pearson frequentist approach has to do with the absence of a reasoned criterion for deciding when an estimated model is adequate on statistical grounds. Goodness-of-fit

criteria have been discredited because of their vulnerability to spurious inference results. Gosset, as the authors rightly observe (Ziliak and McCloskey 2008, 59-60), is credited with raising the issue of invalid probabilistic assumptions, such as 'normality', giving rise to spurious results as early as 1923 (see Lehmann 1999). His questions were explored by Egon Pearson in the early 1930s, but largely ignored by Fisher and the subsequent statistics literature for a variety of reasons beyond the scope of this review.

As argued in Spanos (1986), addressing the problem of *statistical adequacy* (the validation of the model assumptions vis-à-vis data $\mathbf{x}_0$) requires, *ab initio*, a purely probabilistic construal of a statistical model, specified in terms of a complete list of (internally consistent) probabilistic assumptions, in a form that is testable with data $\mathbf{x}_0$. That often requires unveiling implicit assumptions as well as recasting assumptions about unobservable errors terms. It also requires distinguishing between *statistical* and *substantive adequacy*, contrary to the current conventional wisdom in economics which conflates the two under the banner of 'specification error'. This is because securing the former is a necessary condition for assessing the latter (Spanos 2006b). Statistical adequacy renders the relevant error probabilities ascertainable by ensuring that the *nominal* error probabilities for assessing substantive claims are very close to the *actual* ones. The surest way to draw invalid inferences is to apply a 5% significance level test when its actual type I error probability is close to 100% due to misspecification (Spanos and McGuirk 2001).

Using statistical adequacy—not 'oomph' (Ziliak and McCloskey 2008, 48)—to select the best model in the sense that it 'accounts for the regularities in the data', can explain why the t-test, the $R^2$ and other statistics vilified by the authors, are often statistically vacuous when any of the probabilistic assumptions constituting the statistical model in question are invalid for data $\mathbf{x}_0$. Indeed, statistical adequacy helps to place the problem of statistical vs. substantive significance in a proper perspective. Despite the importance of the latter problem, any attempt to address it becomes hopeless unless one deals with the *statistical misspecification* issue first. The very notion of statistical significance becomes ambiguous without statistical adequacy since it is unknown whether the apparent significance is genuine or simply an artifact, i.e., the result of a sizeable discrepancy between the relevant nominal and actual error probabilities; talk about 'baseless size'!

In light of this dubiousness, the researchers accused of 'sizeless stare' and outright ignorance are guilty only of sidestepping a problem which nobody knows how to address adequately, least of all the two authors; paying lip service is far from dealing with it. Continuing this line of reasoning, do the authors expect credit for mentioning a blurred form of the 'specification problem' (Ziliak and McCloskey 2008, xvii) and some vague references to 'other errors', even though they have done nothing about them in their published work? Or do the authors point a finger at the failings of others to distract from the more serious problems that they themselves ignore in their published work?

The problem of statistical misspecification is not only more fundamental, but researchers have known, for some time now, how to handle it using thorough *misspecification testing* and *respecification*. Moreover, Fisher-type significance testing plays a crucial role in model validation (see Spanos 1986, 1999; Mayo and Spanos 2004). Indeed, one wonders how many applied papers published in the *American Economic Review* over the last 30 years are likely to pass the statistical adequacy test; I hazard a guess of less than 1% for the reasons I discuss in Spanos (2006a).

Where does this leave the authors' concern with the problem of statistical vs. substantive significance? Shouldn't they have known that, even if one had a credible procedure to address the problem, one couldn't make any progress on the basis of statistically misspecified models?

In conclusion, do the authors genuinely believe that their 'what to do list', based primarily on (i)-(viii), and some wispy references to "Jeffrey's d, Wald's 'loss function', Savage's 'admissibility' […] and above all Gosset's 'net pecuniary advantage'" (Ziliak and McCloskey 2008, 250), constitute a credible solution to this important problem? If so, they delude themselves far more than those economists at whom they wag their fingers throughout this book.

## REFERENCES

Box, Joan. F. 1978. *R. A. Fisher: the life of a scientist*. New York: Wiley & Sons.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical statistics*. London: Chapman and Hall.

Fisher, Ronald A. 1935. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98 (1): 39-54 [with discussion: 55-82].

Fisher, Ronald A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society*, B, 17 (1): 69-78.

Hacking, Ian. 1965. *Logic of statistical inference*. Cambridge: Cambridge University Press.

Hoover, Kevin D., and Mark V. Siegler. 2008. Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology*, 15 (1): 1–37.

Lehmann, E. L. 1999. Student and small-sample theory. *Statistical Science*, 14 (4): 418-426.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.

Mayo, Deborah G., and Aris Spanos. 2004. Methodology in practice: statistical misspecification testing. *Philosophy of Science*, 71 (5): 1007-1025.

Mayo, Deborah G., and Aris Spanos. 2006. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, 57 (2): 323-357.

Plackett, R. L., and G. A. Barnard (eds.). 1990. *Student: a statistical biography of William Sealy Gosset, based on writings by E. S. Pearson*. Oxford: Clarendon Press.

Savage, Leonard J. 1976. On re-reading R. A. Fisher. *Annals of Statistics*, 4 (3): 441-500.

Spanos, Aris. 1986. *Statistical foundations of econometric modelling*. Cambridge: Cambridge University Press.

Spanos, Aris. 1999. *Probability theory and statistical inference: econometric modeling with observational data*. Cambridge: Cambridge University Press.

Spanos, Aris. 2004. Confidence intervals, consonance intervals, p-value functions and severity evaluations. *Virginia Tech Working paper*, Blacksburg.

Spanos, Aris. 2006a. Econometrics in retrospect and prospect. In *New Palgrave handbook of econometrics, vol. 1*, eds. T. C. Mills, and K. Patterson. London: MacMillan, 3-58.

Spanos, Aris. 2006b. Revisiting the omitted variables argument: substantive vs. statistical adequacy. *Journal of Economic Methodology*, 13 (2): 179-218.

Spanos, Aris, and Anya McGuirk. 2001. The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association*, 83 (5): 1168-1176.

Wald, Abraham. 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10 (4): 299-326.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. Ann Arbor (MI): The University of Michigan Press.

**Aris Spanos** is Wilson Schmidt professor at the Department of economics, Virginia Tech. His research focuses on econometrics, modelling speculative prices, and the philosophy and methodology of empirical modelling. Contact e-mail: <aris@vt.edu>