The Cult of Statistical Significance

By Stephen T. Ziliak and Deirdre N. McCloskey[1]

Roosevelt University and University of Illinois-Chicago

Abstract: 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 (1890-1962), though a genius, was wrong. Fit is not the same thing as importance. Statistical significance is not the same thing as scientific importance or economic sense. But the mistaken equation is made, we find, in 8 or 9 of every 10 articles appearing in the leading journals of science, economics to medicine. The history of this "standard error" of science involves varied characters and plot twists, but especially R. A. Fisher's canonical translation of "Student's" *t*. William S. Gosset aka "Student," who was for most of his life Head Experimental Brewer at Guinness, took an economic approach to the logic of uncertainty. Against Gosset's wishes his friend Fisher erased the consciously economic element, Gosset's "real error." We want to bring it back.

For the past eighty-five years it appears that some of the sciences have made a mistake, by basing decisions on statistical "significance." Though it looks at first like a matter of minor statistical detail, it is not.

Statistics, magnitudes, coefficients are essential scientific tools. No one can credibly doubt that. And mathematical statistics is a glorious social and practical and aesthetic achievement. No one can credibly doubt that, either. Human understanding of chance and uncertainty would be much reduced were it not for Bayes's Theorem, gamma functions, the bell curve, and the rest. From the study of ancient parlor games to the rise of modern space science, mathematical statistics has shown its power.

But one part of mathematical statistics has gone terribly wrong, though mostly unnoticed. The part we are worrying about here seems to have all the quantitative solidity and mathematical shine of the rest. But it also seems—unless we and some other observers of mathematical statistics such as Edgeworth, Gosset, Egon Pearson, Jeffreys, Borel, Neyman, Wald, Wolfowitz, Yule, Deming, Yates, L. J. Savage, de Finetti, Good, Lindley, Feynman, Lehmann, DeGroot, Bernardo, Chernoff, Raiffa, Arrow, Blackwell, Friedman, Mosteller, Kruskal, Mandelbrot, Wallis, Roberts, Granger, Leamer, Press, Moore, Berger, Gigerenzer, Freedman, Rothman, Zellner and a small town of others working in and around the American Statistical Association (see "Works Cited" in Ziliak and McCloskey 2008a, pp. 265-287) are quite mistaken—that reducing the scientific and commercial problems of testing, estimation and interpretation to one of "statistical significance," as some sciences have done for a hundred years, has been an exceptionally bad idea.

Statistical significance is, we argue, a diversion from the proper objects of scientific study. Significance, reduced to its narrow and statistical meaning only—as in "low" observed "standard error" or "$p < .05$"—has little to do with a defensible notion of scientific inference, error analysis, or rational decision making. And yet in daily use it produces unchecked a large net loss for science and society. Its arbitrary, mechanical illogic, though currently sanctioned by science and its bureaucracies of reproduction, is causing a loss of jobs, justice, profit, and even life.

Statistical significance at the 5% or other arbitrary level is neither necessary nor sufficient for proving discovery of a scientific or commercially relevant result. How the

odds should be set depends on the importance of the issues at stake and the cost of getting new material. Let us examine the 5% rule of statistical significance. When a gambler bets at the track for real money, does she insist on 19 to 1 odds (0.95/0.05) before choosing a horse? What does a rational brewer do about the 5% rule when buying hops to make a beer he sells for profit? Should Parliament or Congress enforce a rule of 19 to 1 odds or better for a surgical procedure, newly discovered, which may save the life of the nation's leader? What is scientific knowledge and how does it differ?

We and our small (if distinguished) group of fellow skeptics say that a finding of "statistical" significance, or the lack of it, statistical insignificance, is on its own valueless, a meaningless parlor game. Statistical significance should be a tiny part of an inquiry concerned with the size and importance of relationships. Unhappily it has become a central and standard error of many sciences. The history of this "standard error" of science—the past 85 years of mistaking statistical significance for scientific importance—involves varied characters and plot twists, but especially Ronald A. Fisher's (1890-1962) canonical translation of "Student's" t. William Sealy Gosset (1876-1962) aka "Student," working as Head Experimental Brewer at Guinness's Brewery, took an economic approach to the logic of uncertainty. Fisher erased the consciously economic element, Gosset's "real error." We want to bring it back.

Precision is Nice but Oomph is the Bomb

Suppose you get a call from your Mother, who wants to lose weight. Your Mom is enlisting you—a rational statistician, educator and Web surfer—to find and choose a weight-loss pill. Let's say you do the research and after ranking the various alternatives you deign to propose two diet pills. Mom will choose just one of them. The two pills come at identical prices and side effects (dry mouth, nausea, et cetera) but they differ in weight loss-ability and precision.

The first pill, called "Oomph," will shed from Mom an average of 20 pounds. Fantastic! But Oomph is very uncertain in its effects—at plus or minus 10 pounds (you can if you wish take "plus or minus X-pounds" as a heuristic device to signify in general the amount of "standard error" or deviation around an estimated mean or other coefficient). Speaking concretely, at 20 pounds on average, diet pill Oomph gives Mom a big effect but with a high variance—at plus or minus 10 pounds. Could be ten pounds Mom loses; could be thrice that. After all, some people can and want to lose 30 pounds all at once. (If you doubt this part of the argument just ask your Mom.)

The other pill you found, pill "Precision," will take 5 pounds off Mom on average but it is very precise—at plus or minus 0.5 pounds. Precision is the same as Oomph in price and side effects but Precision is much more certain in its effects. Great! Choosing Precision entails a probable error of plus or minus a mere one-half pound. Its precision is impressive compared to that of pill Oomph, at any rate in view of the design of the experiment that measured the amount of variation in each. Suppose the designs are constant too; allow that the federal government together with the scientific journals has jointly stipulated a statistical protocol (as in fact the U.S. government and many journals do [Ziliak and McCloskey 2008, p. 166]) such that the design and interpretation of experiments on food and drug-related material is more or less the same as to questions of scientific inference—ceteris paribus.

Fine. Now which pill do you choose—Oomph or Precision? Which pill is best for Mom, whose goal is to lose weight?

The problem we are describing is that in the life and human sciences, from agronomy to zoology and including statistics itself, fully 8 or 9 of every 10 publishing scientists chooses Precision over Oomph. From the American Economic Review to the Annals of Internal Medicine the story is the same.

Being precise is not, we repeat, a bad thing.

Consider that statistical significance at some arbitrary level, the favored instrument of precision-lovers, reports on the precision of a particular sort of signal-to-noise ratio, the ratio of the music you can hear clearly relative to the static interference. The signal to noise ratio is analogous to "Student's" ratio. That's a useful ratio, especially so in the rare cases in which the noise of small samples and not of misspecification or other real errors is your chief problem. A high signal-to-noise ratio in the matter of random samples is helpful if your biggest problem is that your sample is too small. Still the signal to noise ratio is by itself a radically incomplete basis for making rational decisions.

The signal to noise ratio is calculated by dividing a measure of what the investigator is curious about—the sound of a Miles Davis number, the losing of body fat, the yield of a barley variety, the impact of the interest rate on capital investment—by a measure of the uncertainty of the signal, such as the variability caused by static interference on the radio or the random variation from a smallish sample. In diet pill terms the noise—the uncertainty of the signal, the variability—is the random effects, such as the way one person reacts to the pill by contrast with the way another person does. In formal hypothesis testing terms, the signal—the observed effect—is typically compared to a "null hypothesis," an alternative belief. The null hypothesis is a belief to test against the data on hand, allowing one to find a difference from it, if there really is one.

In the Precision versus Oomph weight-loss example one can choose the null hypothesis to be a literal zero effect, which is a very common choice of a null (contrast Rothman 1986). That is, the average weight-loss afforded by each diet pill is being tested against the null hypothesis—the alternative belief—that the pill in question will not take off any weight. The magic formula for the signal to noise ratio is:

$$\frac{\text{Observed Effect} - \text{Hypothesized Null Effect}}{\text{Variation of Observed Effect}}$$

Plugging in the numbers from the example yields: (20 – 0)/10 = 2 and (5 – 0)/0.5. = 10. In other words, the signal to noise ratio of pill Oomph is 2-to-1, and of pill Precision 10-to-1. Precision, we find, gives a much clearer signal—five times clearer.

All right, then, once more: which pill for Mother? Recall: the pills are identical in every other way. "Well," say our significance testing colleagues, "the pill with the highest signal to noise ratio is Precision. Precision is what scientists want and what the people, such as your mother, need. So, of course, choose Precision."

But Precision—precision commonly defined as a large t-statistic or small p-value on a coefficient—is obviously the wrong choice. Wrong for Mother's weight-loss plan and wrong for the many other victims of the sizeless scientist. The sizeless scientist decides whether something is important or not—he decides "whether there exists an effect," as he puts it—by looking not at the something's oomph but at how precisely it is estimated. Mom wants to lose weight, not gain precision. Mom cares about the spread around her waist. She cares little—for example, not at all—about the spread around the average of a hypothetically infinitely repeated random sample. The minimax solution (to pick one solution among others in the "loss function" approach of statistics [Press 2003, p. 268]) is obvious: in all states of the world, Oomph dominates Precision. Oomph wins. Choose Oomph. Choosing the inferior pill, pill Precision, though it is highly statistically significant, in fact maximizes *failure*—the ex ante and real failure of Mom to lose up to an additional 24.5 (30-5.5) pounds. You should have picked Oomph.

What the Sizeless Scientists Say in Defense

You will be uneasy. You will say again, "How can this be?" You will suppose that the 5% rule, R. A. Fisher's way, could not be so gravely mistaken—could it? Surely, you will think, a Fisherian disciple of the intellectual quality of Harold Hotelling could not have been confused on the matter of statistical significance. Surely Ziliak and McCloskey and the critics of the technique since the 1880s such as Edgeworth and Gosset and Jeffreys and Deming and Savage and Kruskal and Zellner and Moore and Berger must have it wrong.

Some statisticians and statistical scientists think they disagree with our strictures on null-hypothesis significance testing (Spanos 2008; Ziliak and McCloskey 2008b; Hoover and Siegler 2008; McCloskey and Ziliak 2008). They are made unhappy by our radical assertions. When they do not simply become outraged---which is not uncommon, we said, though never accompanied by actual argument---they cast about for a reply.

Some hope they can trip us up on technical details (we have benefited from comments such as these delivered by Clive Granger, Graham Elliot, Ed Leamer, Joel Horowitz, Arnold Zellner, Jeffrey Wooldridge and others, all published in a special issue of the Journal of Socio-Economics 33 (5), 2004). Technical tripper uppers (including several of our critics in the Journal) observe for example that both diet pills in our Mom example achieved a signal-to-noise ratio of at least 2.0. Therefore, they say, both pills are statistically significant---and so the precision criterion is entirely satisfied! Therefore, they say, the scientist will choose Oomph, which dominates Precision in practical, weight-loss effect. Since the signal-to-noise ratio meets or exceeds 2.0 for each pill, both pills are formally "admissible" (Savage 1954, pp. 114-16). The objection itself illustrates, by the way, the lack of interest in oomph. We stipulated that Mom would choose just one pill. The objectors are not listening to the substance of the science.

But it is anyway irrational, we would reply in technical mode, to make decisions in this sequential (lexicographical) fashion, first precision, then oomph, first step A, then step B, then step C. And even this irrational procedure is not what scientists in the sizeless sciences actually do. They routinely pick the pill with the most precision, not the one with the highest oomph after having established precision. Many articles neglect oomph entirely, assigning to oomph zero weight. They actually never get to step B or C. In radio terms they choose the precise sound of country and western over the crackly sound of Miles Davis on a competing station—not because they like country and western better than jazz but because the signal on the country station is better. (Imagine taking a road trip in the car with Dr. Precision as your companion!) Try to spot precision and oomph in an article funded by the Centers for Disease Control, on drug use among poor women in Atlanta:

> When examined in bivariable analyses, 15 of the 16 temptations-to-use drugs items were found to be associated [that is, statistically significantly related] with actual drug use. These were: while with friends at a party ($p < .001$), while talking and relaxing ($p < .001$), while with a partner or close friend who is using drugs ($p < .001$), while hanging around the neighborhood ($p < .001$), when happy and celebrating ($p < .001$), when seeing someone using and enjoying drugs ($p < .05$), when waking up and facing a tough day ($p < .001$), when extremely anxious and stressed ($p < .001$), when bored ($p < .001$), when frustrated because things are not going one's way ($p < .001$), when there are arguments in one's family ($p < .05$), when in a place where everyone is using drugs ($p < .001$), when one lets down concerns about one's health ($p < .05$), when really missing the drug habit and everything that goes with it ($p < .010$), and while experiencing withdrawal

> symptoms ($p < .01$) . . . . The only item that was not associated with the amount of drugs women used was "when one realized that stopping drugs was extremely difficult"
>
> Journal of Drug Issues 2003, pp. 171-172.

We count 16 instances of precision-only considerations in this one paragraph—and zero instances of oomph. The 2003 referred article about drugs, though funded by the CDC, yields an oomph ratio of 0 percent (total count: 0 oomph, 16 "precision"). The authors believe they are doing serious scientific research, and we suppose that in many ways they are. We find it strange that they used "bivariable" instead of multiple regression techniques. Yet their use of statistical significance is utter nonsense. The whole of their world is significant ($p < .05$). They are in the grips of their own addiction to recreational statistics. Like other sizeless scientists they clearly believe that Fisher's *p*-value gives a quantitative standard of precision though Jeffreys 1961, p. 385 and Zellner 1984, p. 288 and Rothman (1986, chps. 8-10) and others have proven otherwise (Ziliak and McCloskey 2008a, chps. 14-16).

We find the results strange. The part of civilization claiming to set empirical standards for science and policy has decided to use illogical instruments and irrelevant empirical standards for science and policy. In journals such as Nature, Science, The New England Journal of Medicine, The Journal of Clinical Psychiatry, Annals of Internal Medicine, Educational and Psychological Measurement, Epidemiology and Infection, Administrative Science Quarterly, Decision Sciences, and the American Economic Review, size doesn't matter. Oomph doesn't matter. Something is wrong.

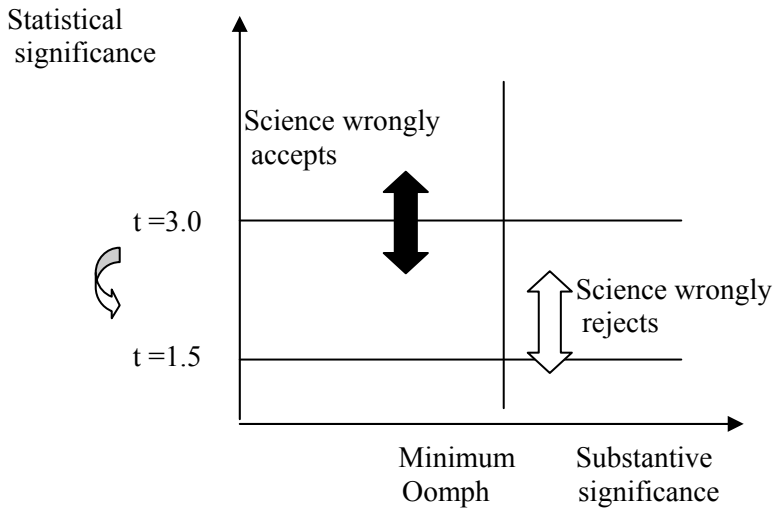
Significant Does Not Mean Important and Insignificant Does Not Mean Unimportant

"Somehow the numbers are rigged," a few critics say, "to get the results you want." More technical fixes. No. We can change the numbers in the example, leaning harder against Oomph, if you want, and yet come to the same conclusion: that focusing on Precision alone is mistaken. Leave the pill Precision at "takes off on average 5 pounds plus or minus 0.5 pounds," but change the statistics for the Oomph pill: suppose it takes off 20 pounds on average but now it is even more uncertain in its effects. Before, it varied by plus or minus 10 pounds. Now suppose Oomph varies by plus or minus 14 pounds.

The signal-to-noise ratio for pill Precision is still 10.0 to 1.00—very precise. Pill Oomph, however, is now much less precise, only 1.43 to 1.00. That's noisy data by Fisher's arbitrary Rule of Two. It yields a very imprecisely estimated average for Oomph's effect, at any rate if your notion of "precision" is "sampling odds alone." At 1.43 the signal-to-noise ratio is a good deal less than 1.96. Under the convention of 5% significance, it is statistically insignificant"—ineligible for publication and government approval.

But, we ask again, which pill for Mother, whose goal in all this is to lose weight, not to obey a rule-bound philosophy of science? Will Oomph promises to shed between 6 and 34 pounds while Precision's lower and upper bound are 4.5 and 5.5 pounds. Oomph, despite its lack of precision, is still what Mother wants (Figure 1).

Figure 1
Size Matters No Matter Where You Draw the Line of Precision



Source: Ziliak and McCloskey 2008, Figure 3.1, p. 44

The precision-only statistician might say, "Wait a minute. How much weight does Mother want to lose? If 34 pounds of loss would leave her a skeleton. . . ." But here the precision-only statistician is clearly conceding our case.

"Significance" and Merck

Sadly, such basic size-matters/how-much questions are not being asked or answered, not even (it would seem) by government regulators. Take the case of a pain relief pill, such as Vioxx and its side effects. How much Vioxx is too much for Mom's fragile heart? Merck, the producer of Vioxx, did not say – or would not say until the medics and lawyers showed up. Yet Merck was in 2005 the third-largest drug manufacturer in the United States. Its painkiller Vioxx was first distributed in the United States in 1999, and by 2003 had been marketed in over 80 different countries. At its peak in 2003 Vioxx (also called Ceoxx) brought in some $2.5 billion. In that year a 73-year old woman died suddenly of a heart attack while taking her prescribed Vioxx pills. Anticipating a law suit the senior scientists and company officials at Merck, newspaper accounts have said, huddled over the statistical significance of the original clinical trial.

From what an outsider can infer the report of the clinical trial appears to have been fudged. Data that made Vioxx look bad were allegedly simply omitted from the report. A rheumatologist at the University of Arizona and lead author of the 2003 Vioxx study, Jeffrey Lisse, admitted later that not he but "Merck actually wrote the report." Perhaps there is some explanation consistent with a more reputable activity than data fudging. We don't know.

"Data fudging and significance testing are not the same," you will say. "Most of us do not commit fraud." True. We agree. But please listen.

The clinical trial was conducted in 2000 and the findings were published three years later in the Annals of Internal Medicine (Lisse et al. 2003). The scientific article reported "that five [note the number, 5] patients taking Vioxx had suffered heart attacks during the trial, compared with one [note the number, 1] taking naproxen [the generic

drug, such as Aleve, given to a control group], a difference that did not reach statistical significance."[2] The signal-to-noise ratio did not rise to 1.96, the 5% level of significance that the Annals of Internal Medicine uses as strict line of demarcation, discriminating the "significant" from the insignificant, the scientific from the non-scientific.

Therefore, Merck claimed, there was no difference in the effects of the two pills. No difference in oomph, they said, despite a Vioxx disadvantage of about 5-to-1. Then the apparent fraud: the article neglected to mention that in the same clinical trial three additional takers of Vioxx, including the 73-year old woman whose survivors brought the problem to public attention, suffered from heart attacks. Eight in fact suffered or died in the clinical trial, not five. It appears that the scientists, or the Merck employees who wrote the report, simply dropped the three observations.

Why? Why did they drop the 3? We do not know for sure. The courts will decide. But an outsider could be forgiven for inferring that they dropped the three observations in order to get an amount of statistical significance low enough to claim—illogically, but this is the usual procedure—a zero effect.

In this case as in many other cases uncomfortably familiar to the variable dropper-and-adder, the reasoning is that if you can keep your sample small enough—by dropping parts of it, for example—you can claim "insignificance," and continue marketing. In the published article on Vioxx you can see that the authors believed they were testing, with that magic formula, that sizeless stare, whether an effect "existed." "The Fisher exact test," they wrote in typical sizeless scientific fashion, and in apparent ignorance of the scientific values of his nemesis, W.S. Gosset, "was used to compare incidence of confirmed perforations, ulcers, bleeding, thrombotic events, and cardiovascular events" (Lisse et al. 2003, p. 541). "All statistical tests . . . were performed at an α level of 0.05" (p. 541).

If the Merck scientists could get the number of heart attacks down to 5, you see, they could claim to other sizeless scientists that the harmful effect wasn't there, didn't exist, had no oomph, was in fact zero. The damage was actually: naproxen takers, 1 victim, Vioxx takers, 8 victims, not 5. Other things equal, the relative toll of Vioxx to naproxen was 8-to-1, leaning strongly against Vioxx. And with the sample size the scientists had the true 8 heart attacks were in fact statistically significant even by the 5% Fisher criterion—good enough, that is, on grounds of sampling precision, to be counted as a "finding" in the Annals. But Merck didn't want to find that their Vioxx was dangerous. So they pretended that the deaths were insignificant.

In a scientific culture depending on a crude version of precision and the sizeless stare, "significance" was sociologically speaking Merck's problem. Merck wanted the unfavorable results to be statistically insignificantly different from a zero effect, in order that it could claim no effect. Merck, while earning billions, did not have a grip on its loss function; but it understood the cultural and legal significance of "significance"—perhaps a little too well.

How the Journals, Textbooks, and Grantors Enforce the Sizeless Stare

The proximate cause of the unhappy situation is similar across other disciplines: students get bad advice from teachers and then they grow up to be scientists, judges, and teachers themselves. In economics departments almost all of the teachers of probability, statistics and econometrics claim that statistical significance is the same thing as scientific significance. An econometrician at Oxford University, our friend David Hendry, for example, is famous for saying "test, test, test," where the phrase means "Fisher, Fisher, Fisher". Most statistical textbooks in any field, from theoretical statistics

down to the merest cookbook, recommend the same (for example Spanos (1986), authored by a former Hendry student, does.)

A few get it right. The late Morris DeGroot, a Fellow of the American Statistical Association and Fellow of a half dozen other scientific associations, not to mention a distinguished statistical theorist and teacher of several Nobel laureates in economics at Carnegie-Mellon University, wrote as follows in his exemplary textbook of 1975:

> It is extremely important. . . to distinguish between an observed value of U that is statistically significant and an actual value of the parameter. . . . In a given problem, the tail area corresponding to the observed value of U might be very small; and yet the actual value . . . might be so close to [the null] that, for practical purposes, the experimenter would not regard [it] as being [substantively] different from [the null].
> DeGroot 1975, p. 496.

DeGroot does not leave the matter as a throw-away point, a single sentence in an otherwise Fisherian tract, as so many of even the minority of statistics books that mention the matter do. On the contrary, DeGroot goes on: "It is very likely that the t-test based on the sample of 20,000 will lead to a statistically significant value of U. . . . [The experimenter] knows in advance that there is a high probability of rejecting [the null] even when the true value . . . differs only slightly from [the null]" (DeGroot, p. 497).

But few econometrics textbooks make the distinction between statistical and economic significance. Even the best books do not give anything close to equal emphasis on economic significance—not in hundreds, of pages devoted to explaining Fisherian significance. In the texts widely used in the 1970s and 1980s, for example, when the bad practice was becoming standard, such as Jan Kmenta's Elements of Econometrics (1971) or John Johnston's various editions of Econometric Methods (1963, 1972, 1984) there are no mentions of economic significance. Peter Kennedy, in his A Guide to Econometrics (1985), briefly mentions that a large sample always gives "significance." This is part of the point, but not nearly all of it, and in any case it is relegated to an endnote (p. 62). Kennedy says nothing else on the matter, as if economic statistics is a branch of vague mathematics.

Clive Granger on Not Mentioning Economic Significance

Arthur Goldberger gives the topic of "Statistical vs. Economic Significance" a page of his A Course in Econometrics (1991), quoting a little article by McCloskey in 1985. Goldberger's lone page has been flagged as unusual. The Nobel laureate Clive Granger reviewed four econometrics books in the March 1994 issue of the Journal of Economic Literature and wrote: "when the link is made [in Goldberger between economic science and the technical statistics] some important insights arise, as for example the section [well . . . the page] discussing 'statistical and economic significance,' a topic not mentioned in the other books" [by R. Davidson and J. G. MacKinnon, W. H. Greene, and W. E. Griffiths, R. C. Hill, and G. G. Judge] (Granger 1994, p. 118, italics supplied).

Not mentioned. That is the standard for education in econometrics and statistics at the advanced level. The three stout volumes of the Handbook of Econometrics contain a lone mention of the point, unsurprisingly by Edward Leamer.[3] In the 732 pages of the Handbook of Statistics there is one sentence (p. 321, Florens and Mouchart, in Maddala, Rao, and Vinod, eds., 1993). Aris Spanos has in his impressive Probability Theory and Statistical Inference (1999) tried to crack the Fisher monopoly on advanced econometrics, but even Spanos looks at the world with a sizeless stare (pp. 681-728). His history of hypothesis testing has in any case been ignored.

In the elementary courses is the elementary point is made? Takeshi Amemiya's advanced textbook in econometrics (1985), for instance, which for a while marked the peak of accomplishment in such matters, never distinguishes economic from statistical significance, nowhere in hundreds of pages.

The Stanford University professor of econometrics recommends that the student prepare for his highly mathematical book "at the level of Johnston, 1972." Examine Jack Johnston's book, then, as Amemiya recommends, in Johnston's most comprehensive edition (Johnston 1984). Johnston uses the term "economic significance" once only, rather late in his formerly popular book, while discussing a technique commonly used outside of economics, without contrasting it with statistical significance, on which he has lavished by then hundreds of pages: "It is even more difficult to attach economic significance to the linear combinations arising in canonical correlation analysis than it is to principal components" (p. 333; italics supplied).

At the outset, in an extended example of hypothesis testing spanning pages 17 to 43, he goes about testing in the orthodox Fisherian way. In a toy example he tests the hypothesis that "sterner penalties" for dangerous driving in the United Kingdom would reduce road deaths, and concludes that "the computed value [of the t-statistic] is suggestive of a reduction, being significant at the 5 percent, but not at the one percent level" (p. 43). What does this mean? Johnston suggests that at a more rigorous level—1 percent—you might not act on the result, although acting on it would have saved about 100,000 lives in the UK over the period 1947-57. Johnston has merged statistical and policy significance. Sterner penalties, according to his data, save lives. The rigorously 1-percent statistician, Johnston implies, would ignore this fact. By what warrant?

A tenacious defender of the prevailing method might argue that Johnston in turn had assumed that his readers got their common sense from still more elementary courses and books. Johnston directs the reader who has difficulty with his first chapter to a "good elementary book" (p. ix), mentioning Hoel's Introduction to Mathematical Statistics (1954), Mood's Introduction to the Theory of Statistics (1950), and Fraser's Statistics: An Introduction (1958). These are fine books. Mood's book gives a treatment of power functions, for example, which a modern economist would do well to read.

But none of the three books makes a distinction between substantive and statistical significance. Hoel for example writes:

> There are several words and phrases used in connection with testing hypotheses that should be brought to the attention of students. When a test of a hypothesis produces a sample value falling in the critical region of the test, the result is said to be significant; otherwise one says that the result is not significant.
>
> Hoel 1954, p. 176, his italics.

R.-A.-Fisher Significance. That is all.

The old classic by W. Allen Wallis and Harry Roberts, Statistics: A New Approach, first published in 1956, is an exception:

> It is essential not to confuse the statistical usage of "significance" with the everyday usage. In everyday usage, "significant" means "of practical importance," or simply "important."
>
> Wallis and Roberts 1965, p. 385, italics supplied.

The point has been revived in some elementary statistics books, though not in most of them. Older books tend to be better. In their leading textbook the statisticians Freedman, Pisani, and Purves (1978, p. 487; compare pp. 501, A-23) could not be plainer. The distinction is emphasized in the elementary books by

Wonnacott and Wonnacott (1982, p. 160; one of the brothers is an economist, the other is a statistician) and David S. Moore and George McCabe (1993, p. 474) are very clear. (David Moore, a past-president of the American Statistical Association, is like his fellow past-presidents of the ASA—ignorable people such as Shewhart, Kruskal, Wallis, Mosteller, Zellner—*strongly against the conventional culture of significance testing.*) Lately in econometrics Jeffrey Wooldridge (2000) is a standout, giving about 3 pages to the difference between economic and statistical significance. But 3 pages out of several hundreds? Is less than 1% of total output the right oomph-to-precision ratio for science?

Statistical Education in Psychology and Related Areas

The precision-only problem is worse in psychology and other rule-bound fields, guided by manuals. The history of the Publication Manual of the American Psychological Association exhibits the significance problem for more fields than psychology alone. The Manual sets the editorial standards for over a thousand journals in psychology, education, and related disciplines, including forensics, social work, and parts of psychiatry.

In the 1952 first edition of the Manual the thinking was thoroughly pro-Fisher and anti-Gosset, obsessed with significance: "Extensive tables of non-significant results are seldom required," it says. "For example, if only 2 of 20 correlations are significantly different from zero, the two significant correlations may be mentioned in the text, and the rest dismissed with a few words"[4] The Manual was conveying what Fisher and Hotelling and others, such as Klein in economics and A. W. Melton in psychology, were preaching at the time. In the second edition—twenty years on—the obsession became compulsion:

> Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more
>
> APA 1974, p. 19; quoted in Gigerenzer 2004, p. 589.

Recent editions of the Manual—as both critics and defenders of the Establishment observe— do at last recommend that the authors report "effect size."[5] The fifth edition, published in 2001, added exact levels of significance to analysis of variance. But as Gerd Gigerenzer (2004), a leading student of such matters, observes, the Manual retained also the magical incantations of $p < .05$ and $p < .01$. Bruce Thompson, a voice for oomph in education, psychology, and medicine, commends the fifth edition for suggesting that confidence intervals are the "best reporting strategy."[6] Yet, as Thompson and Gigerenzer and Fiona Fidler's team of researchers have noted, in Gigerenzer's words, "The [fifth] manual offers no explanation as to why both [confidence intervals for effect size and asterisk-superscripted p-values] are necessary . . . and what they mean" (Gigerenzer, p. 594). The Manual offers no explanation for the significance rituals—no justification, just a rule of washing one's hands of the matter if $p < .05$ or $t > 2.00$.

In psychology and related fields the reforms of the 1990s were nice sounding but in practice ineffectual. The 2001 edition of the Manual appears to reflect pressure exerted by editors and scientists intent on keeping their machine for article-producing well oiled. Some 23 journals in psychology and education now warn readers and authors against the sizeless stare.[7] It is about 2% of the journals.

In 1950 A. W. Melton assumed editorship of the trend-setting Journal of Experimental Psychology. In 1962 Melton described what had been his policy for accepting manuscripts at the journal (Melton 1962, pp. 553-7). An article was unlikely to be published in his journal, Melton said, if it did not provide a test of significance and in particular if it did

not show statistically significant results of the Fisher type. Significance at the 5% level was "barely acceptable"; significance at the 1% or "better" level was considered "highly acceptable," and definitely worthy of publication (p. 544). Melton justified the rule by claiming that it assured that "the results of the experiment would be repeatable under the conditions described." Gigerenzer et al. note that after Melton's editorship it became virtually impossible to publish articles on empirical psychology in any subfield without "highly" statistically significant results.

In a penetrating article of 1959, "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa," the psychologist Thomas D. Sterling surveyed 362 articles published in four leading journals of psychology: Experimental Psychology (Melton's journal), Comparative and Physical Psychology, Clinical Psychology, and Social Psychology, testing Melton's claims (Sterling 1959). Sterling was not surprised when he found that only 8 of 294 articles published in the journals and using a test of significance failed to reject the null. Nearly 80% of the papers relied on significance tests of the Fisherian type to make a decision (286 of 362 published articles). And, though Sterling does not say so, every article using a test of significance—that is, those 80% of all the articles—employed Fisher's 5% philosophy exclusively. (Melton's stricter rule of 1% was adopted by some of the journals.) The result "shows that for psychological journals a policy exists under which the vast majority of published articles satisfy a minimum criterion of significance" (Sterling 1959, p. 31).

Sterling observed further that despite a rhetoric of validation through replication of experiments---to which Gosset (Ziliak 2008) gave much of his scientific life, by the way, quite unlike Fisher, who preferred to do more statistical calculations on existing data---not one of the 362 research articles was a replication of previously published research. From his data Sterling derived two propositions:

> A1: Experimental results will be printed with a greater probability if the relevant test of significance rejects H_0 for the major hypothesis with $\Pr(E \mid H_0) \leq .05$ than if they fail to reject at that level.
>
> A2: The probability that an experimental design will be replicated becomes very small once such an experiment appears in print.
>
> Sterling 1959, p. 33.

He understated. Nearly certainly an experimental result that "fails to reject" will not be printed, and by A. W. Melton with probability 1.0. And why actually replicate when the logic of Fisherian procedures gives you an imaginary replication without the bother and expense?

"A picture emerges," wrote Sterling with gentle irony, "for which the number of possible replications of a test between experimental variates is related inversely to the actual magnitude of the differences between their effects. The smaller the difference the larger may be the likelihood of repetition" (p. 33).

The Sociology of Ambition: The Fable of Gosset the Bee and Fisher the Wasp

But of course the journals and textbooks were framed by forces bigger than them, and Fisher came out hard. In the book we argue that a number of different sociological, political and socio-economic forces converge to explain the origin, institutionalization, and 85-year long persistence of "the standard error" (Ziliak and McCloskey 2008, chp. 24). Merton's "bureaucratization of knowledge," Hayek's "scientism," and High Modernism, we argue are three of the biggies. But the origin and persistence of the problem we have traced largely to a world-famous man, his anonymous friend, and a

little-known series of debates between them. R. A. Fisher may have campaigned for the mistaken use of "Student's" t but he did not learn his views from "Student." W. S. Gosset aka "Student" was, we have mentioned, Fisher's behind the scene teacher and friend (Ziliak 2008). Here are representative instructions from Fisher on how to interpret "significance":

> It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . *Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance*
>
> Fisher 1926b, p. 504, italics supplied.

> It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to *eliminate from further discussion* the greater part of the fluctuations which chance causes have introduced into their experimental results
>
> Fisher 1935 [1960], p. 13, italics supplied.

To Fisher a statistically "significant" result is the finding; no evaluation of a coefficient, mean difference or model is necessary or possible in his theory:

> Finally, in inductive inference we introduce no cost functions for faulty judgments . . . In fact, scientific research is not geared to maximize the profits of any particular organization, but is rather an attempt to improve public knowledge undertaken as an act of faith to the effect that, as more becomes known, or more surely known, the intelligent pursuit of a great variety of aims, by a great variety of men, and groups of men, will be facilitated. We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any currency.
>
> Fisher 1955, p. 75, italics supplied.

Contrast Fisher's "faith" in the 5% rule, his anti-evaluation evaluation of "significance," with Gosset's real economic approach. Gosset himself, the eldest son of a Royal Combat Engineer, never believed "significance" that could be a substitute for finding out How Much. He was one of nature's economists, and was required to act as a profit center at Guinness, where he was Apprentice Experimental Brewer (1899-1906), Head Experimental Brewer (1907-1935), and for the rest of his short life Head Brewer at Park Royal and Dublin (1935-1937). Evidence for or against a hypothesis was, Gosset told Karl Pearson in a letter of 1905, a matter of net "pecuniary advantage:"

> My original question and its modified form. When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery"], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient

> in such work as ours and I advised that some outside authority [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment. This is one of the points on which I should like advice.
>
> Gosset, c. April 1905, in E. S. Pearson 1939, pp. 215-216; italics supplied.

Pearson didn't understand the advice. The great man of large samples never did grasp Gosset's point--though wisely he agreed to publish "Student's" papers (Student 1942).

Gosset seems never to have tired of teaching it. Twenty years after his letter of 1905 he responded to a query by Egon Pearson (1895-1980), eldest son of the great Karl, who unlike Pearson père definitely did grasp the point. Gosset in letters to Egon improved on his already sound definition of substantive significance. To net pecuniary value he added that before she can conclude anything decisive about any particular hypothesis the statistician must account for the expected "loss" [in lives or pounds sterling] relative to some "alternative hypothesis."[8] Gosset explained to Pearson fils that the confidence we place on or against a hypothesis depends entirely on the confidence and real-world relevance we put on some other hypothesis, possibly more true and important. Egon, with Jerzy Neyman, would later call Gosset's idea "power." Egon never failed to credit Gosset for inspiring the core idea.

In 1937 Gosset, the inventor and original calculator of "Student's" t-table told Egon, then editor of *Biometrika*, that a significant finding is by itself "nearly valueless":

> obviously the important thing in such is to have a low real error, not to have a "significant" result at a particular station. The latter seems to me to be nearly valueless in itself. . . . Experiments at a single station [that is, tests of statistical significance on a single set of data] are almost valueless. . . . What you really want is a low real error. You want to be able to say not only "We have significant evidence that if farmers in general do this they will make money by it", but also "we have found it so in nineteen cases out of twenty and we are finding out why it doesn't work in the twentieth." To do that you have to be as sure as possible which is the 20th—your real error must be small
>
> Gosset to E. S. Pearson 1937, in Pearson 1939, p. 244.

Gosset, we have noted, is unknown to most users of statistics, including economists. Yet he was proposing and using in his own work at Guinness a characteristically economic way of looking at the acquisition of knowledge and the meaning of "error." The inventor of small sample econometrics focused on the opportunity cost of each observation; he tried to minimize random and non-random errors, real errors. (The quality of his results—the power of his tests—you may sample for yourself—with fish n'chips or a la cart.) Gosset's way at Guinness became the way of Neyman and Pearson, of Wald and Savage, and of Jeffreys and Zellner and Lehman and Press and Berger, a loss function approach which has been all but crowded out of science by the will of an ingenious arbitrary ruler's illogical campaign. The cult continues to pick the wrong pills and policies and commercial projects. It makes no sense. As Deming (1975, p. 152) put it long ago: "Small wonder that students have trouble [learning significance testing]. They may be trying to think."

Select References

De Finetti, Bruno. 1971 [1976]. "Comments on Savage's "On Rereading R. A. Fisher." Annals of Statistics 4(3): 486-7.

DeGroot, Morris H. 1975 [1989]. Probability and Statistics. Reading, MA: Addison-Wesley.

Deming, W. Edwards. 1975. "On Probability as a Basis for Action," American Statistician 29 (4): 146-52.

Fisher, R. A. 1925a[1941]. Statistical Methods for Research Workers. New York: G. E. Stechart and Co.

Fisher, R. A. 1925b. "Applications of 'Student's' distribution." Metron V(3, Dec.): 90-104.

Fisher, R. A. 1925c. "Expansion of 'Student's' Integral in Powers of n^{-1}." Metron V(3, Dec.): 110-120.

Fisher, R. A. 1926a. "Arrangement of Field Experiments." Journal of Ministry of Agriculture XXXIII: 503-13.

Fisher, R. A. 1935. The Design of Experiments. Edinburgh: Oliver & Boyd. Reprinted in eight editions and at least four languages.

Fisher, R. A. 1955. "Statistical Methods and Scientific Induction." Journal of the Royal Statistical Society, Series B (Methodological), Vol. 17, No. 1, pp. 69-78.

Fisher, R. A. 1956 [1959]. Statistical Methods and Scientific Inference. New York: Hafner. Second edition.

Elliott, Graham, and Clive W. J. Granger. "Evaluating Significance: Comments on 'Size Matters.'" Journal of Socio-Economics 33(5): 547-550.

Gigerenzer, Gerd. 2004. "Mindless Statistics." Journal of Socio-Economics 33(5): 587-606.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. 1989. The Empire of Chance. Cambridge: Cambridge University Press.

Gosset, William S. [aka "Student"] 1904. "The Application of the 'Law of Error' to the Work of the Brewery," Laboratory Report, Vol. 8, Nov. 3, Arthur Guinness & Son, Ltd., pp. 3–16 and unnumbered appendix (Guinness Archives, Diageo).

Hoover, Kevin and Mark Siegler. 2008. "Sound and Fury: McCloskey and Significance Testing in Economics," Journal of Economic Methodology 15 (no. 1, March): 1-37.

Jeffreys, Harold. 1961 [1939]. Theory of Probability (Oxford Classics). Third edition.

Kruskal, William H. 1980. "The Significance of Fisher: A Review of R. A. Fisher: The Life of a Scientist, Journal of the American Statistical Association, Vol. 75, No. 372 (Dec.):1019-30.

McCloskey, Deirdre N. and Stephen T. Ziliak. 2008. "Signifying Nothing: Reply to Hoover and Siegler," Journal of Economic Methodology 15 (no. 1, March): 39-55

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. "The Standard Error of Regressions," Journal of Economic Literature 34 (March 1996): pp. 97-114.

Moore, David S., and George P. McCabe. 1993. Introduction to the Practice of Statistics. New York: Freeman.

Pearson, Egon S. 1990 [posthumous]. 'Student': A Statistical Biography of William Sealy Gosset. Oxford: Clarendon Press. Edited and augmented by R. L. Plackett, with the assistance of G. A. Barnard.

Pearson, Egon S. 1939. "'Student' as Statistician." Biometrika 30 (3/4, Jan.): 210-50.

Press, S. James. 2003. Subjective and Objective Bayesian Statistics. New York: Wiley.
Rothman, Kenneth J. 1986. Modern Epidemiology. New York: Little, Brown.
Savage, L. J. 1954. Foundations of Statistics. New York: Dover.
Spanos, Aris. 2008. "Review of Stephen T. Ziliak and Deirdre N. McCloskey's The Cult of Statistical Significance [. . .]," Erasmus Journal for Philosophy and Economics I (no. 1, Autumn): 154-64.
Spanos, Aris. 1986. Statistical Foundations of Econometric Modeling. Cambridge: Cambridge University Press.
Student. 1942 [posthumous]. Student's Collected Papers (London: Biometrika Office). Eds. E. S. Pearson and John Wishart.
Student. 1908a. "The Probable Error of a Mean." Biometrika VI (1, March): pp. 1-24.
Student. 1908b. "The Probable Error of a Correlation Coefficient." Biometrika (2/3, Oct.): pp. 300-310.
Student. 1923. "On Testing Varieties of Cereals." Biometrika 15 (3/4, Dec.): 271-293.
Student. 1925. "New Tables for Testing the Significance of Observations." Metron V(3, Dec.): 105-108.
Zellner, Arnold. 2005. Statistics, Econometrics, and Forecasting (Cambridge: Cambridge University Press). The Stone Lectures in Economics.
Zellner, Arnold. 2004a. "To Test or Not to Test and If So, How?: Comments on 'Size Matters,'" Journal of Socio-Economics 33(5): 581-86.
Zellner, Arnold. 1984. Basic Issues in Econometrics. Chicago: Univ. of Chicago Press.
Ziliak, Stephen T. 2008. "Guinnessometrics: The Economic Foundation of "Student's" t," Journal of Economic Perspectives 22(4, Fall): 199–216.
Ziliak, Stephen T., and Deirdre N. McCloskey. 2008a. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives (Ann Arbor: University of Michigan Press).
Ziliak, Stephen T., and Deirdre N. McCloskey. 2008b. "Science is judgment, not only calculation: A Reply to Aris Spanos's review of The Cult of Statistical Significance," Erasmus Journal for Philosophy and Economics I (no. 1, Autumn): 165-70.
Ziliak, Stephen T., and Deirdre N. McCloskey. 2004a. "Size Matters: The Standard Error of Regressions in the American Economic Review." Journal of Socio-Economics 33(5): 527-46.
Ziliak, Stephen T., and Deirdre N. McCloskey. 2004b. "Significance Redux." Journal of Socio-Economics 33(5): 665-75. Replies to comments by Elliot, Granger, Horowitz, Leamer, O'Brien, Thorbecke, and Zellner.

---

[1] "The Cult of Statistical Significance" was presented at the Joint Statistical Meetings, Washington, DC, August 3rd, 2009, in a contributed session of the Section on Statistical Education. For comments Ziliak thanks many individuals, but especially Sharon Begley, Ronald Gauch, Rebecca Goldin, Danny Kaplan, Jacques Kibambe Ngoie, Sid Schwartz, Tom Siegfried, Arnold Zellner and above all Milo Schield for organizing an eyebrow-raising and standing-room only session.

[2] New York Times News Service, Alex Berenson, Chicago Tribune, April 24, 2005, Sect. 1, p. 14; italics supplied.

[3] Leamer 1983, p. 325, in Griliches and Intriligator, eds., 1983, Vol. I.

[4] APA Manual 1952, p. 414; quoted in Fidler et al. 2004, p. 619.

[5] APA Manual 1994, p. 18; Thompson 2004, p. 608; Fidler et al., 2004, p. 619.

[6] APA Manual 2001, p. 22; Thompson 2004, p. 609.

[7] Hill and Thompson 2004, in Fidler 2004, p. 619.

[8] Letter no. 1 of Gosset to E. S. Pearson, May 11, 1926, quoted in Pearson 1939, p. 243.