The Standard Error:

How Some Sciences Lost Interest in Magnitude, and What to Do About It

By

Stephen T. Ziliak

and

Deirdre N. McCloskey

Preface

The implied reader of the book is a significance tester, the keeper of numerical things. 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, though a genius, was wrong. Fit is not

the same thing as importance. Statistical significance is not the same thing as scientific and human relevance. R^2, *t*-statistic, *F*-test, "exact *p* value," and all the more sophisticated versions of them in time series and the most advanced statistics are misleading, at best.

No working scientist today knows much about Gosset, a master brewer of Guinness stout and the inventor of a good deal of modern statistics. The scruffy little Gosset, with his tall leather boots and a rucksack on his back, is the heroic underdog in our story. Gosset, we claim, was a great scientist. He took an economic approach to the logic of uncertainty. He invented the economic and statistical design of experiments. For over two decades he quietly tried to educate Fisher, who was always eager for Gosset's advice. But Fisher, our flawed villain, erased from Gosset's inventions the consciously economic element. We want to bring it back.

We lament what could have been in statistical sciences if only Fisher had cared to understand the full import of Gosset's insights. Or if only Egon Pearson had had the forceful personality of his father, Karl. Or if only Gosset had been a professor and not a business man, and had been positioned to offset the intellectual capital of Fisher.

But we don't consider the great if mistaken Fisher and his intellectual descendents our enemies. We have learned a great deal from Fisher and his followers, and still do, as many have. We hope you, a significance tester, will read our book optimistically---with a sense of how "real" significance can transform your science. Biometricians who study AIDS and economists who study growth policy in poor countries are causing damage with a broken statistical instrument. But wait: consider the progress we can make if we fix the instrument.

Can so many scientists have been wrong over the eighty years since 1925? Unhappily, yes. The mainstream in science, as any scientist will tell you, is often wrong. Otherwise, come to think of it, science would be complete. Few scientists would make that claim. Statistical significance is surely

not the only error in modern science, though it has been we will show an exceptionally damaging one. Scientists are often tardy in fixing basic flaws in their sciences, despite the presence of better alternatives. Think of the half century it took American geologists to recognize the truth of drifting continents, a theory proposed in 1915 by---of all eminently ignorable people---a German meteorologist. Change takes time. Scientists, after all, are human. What Nietzsche called the "twilight of the idols," the fear of losing a powerful symbol or god or technology, haunts us all.

In the statistical fields such as economics, psychology, sociology, medicine the idol is the test of significance. The alternative, Gossetian way is a uniformly more powerful test, but has been largely ignored. Unlike the Fisherian idol, Gosset's approach is a rational guide for decision-making, and easy to understand. But it has been resisted now—flat ignored--for eighty years.

Our book also addresses implied readers outside the statistical fields themselves, such as intellectual historians and philosophers of science. The history and philosophy of applied statistics took a wrong turn in the 1920s. In an admittedly sketchy way---Ziliak himself is working on a book centered on Gosset---we explore the philosophy and tell the history here. We found that the recent historians of statistics, whom we honor in other matters, have not gotten around to Gosset. The historiography of "significance" is still being importantly shaped by R. A. Fisher himself, four decades beyond the grave. It is known among sophisticates that Fisher took pains to historicize his prejudices about statistical methods. Yet his history gave little credit to other people, and none to those who in the 1920s developed a decision-theoretic alternative to the Fisherian routine. Since the 1940s most statistical theorists, particularly at the advanced level, do not mention Gosset. With the notable exception of Donald MacKenzie, a sociologist and historian of science,

scholars have not much examined Gosset's published works. And it appears that no one besides the ever-careful Egon S. Pearson (1895-1980) has looked very far into the Gosset archives—and that was in 1937-1939, for the purpose of an obituary.

The evidence on the Gosset-Fisher relationship that Ziliak found in the archives is startling. In brief, Gosset got scooped. Fisher's victory over Gosset has been so successful and yet so invisible that a 2006 publication of *anti*-Fisherian statistics makes the usual mistake, effectively equating Fisher's approach with that of Gosset's (Howson and Urbach 2006, p. 133). In truth it was Gosset, in 1905, not Neyman, in 1938, who gave "the first emphasis of the behavioralistic outlook in statistics" (Savage 1954, p. 159).

Only slowly did we realize how widespread the standard error had become in sciences other than our home field of economics. Some time passed before we systematically looked into them. Thus the broader intervention here. We couldn't look into to every science or subfield. And additional work remains of course to be done on significance and other problems of testing and estimation. Some readers, for example, have asked us to wade in on the dual problems of specification error and causality. We reply that we agree---these are important issues---but we couldn't do justice to them here.

But we think the methodological overlaps in education and psychology, economics and sociology, agriculture and biology, economics and epidemiology are sufficiently large, and the inheritance in them of Fisherian methods sufficiently deep, that our book can shed light on all the *t*-testing sciences. We were alarmed and dismayed to discover, for example, that supreme courts in the U. S., state and Federal, have been deciding cases on the basis of Fisher's arbitrary test. The law is becoming infected with Fisher. Time to speak up.

We invite also a general and non-technical reader to the discussion, too. If he starts at the

beginning and reads through Chapter 3 he will get the main point—that oomph, the difference a treatment makes, dominates precision. The extended but simple "diet pill example" in Chapter 3 will equip him with the essential logic, and with the replies he'll need to stay in the conversation. Chapter 17 through to the end of the book provides our brief history of the problem, and a sketch of a solution.

Readers may find it strange that two historical economists such as we have intruded on the theory, history, philosophy, sociology, and practice of hypothesis testing in the sciences. We are not professional statisticians, and are only amateur historians and philosophers of science. Yet economists have played a role in the logic, philosophy, and dissemination of testing, estimation, and error analysis in all of the sciences, from Bernoulli to Mill to Friedman to Heckman. Gosset himself, we've noted, was a business man, and the inventor of the economic approach to uncertainty. Keynes wrote *A Treatise on Probability* (1921), an important if somewhat neglected book on the history and foundations of probability theory.

Advanced empirical economics, which we've endured, taught, and written about for years, has become an exercise in hypothesis testing, and is broken. We're saying here that the brokenness extends to many other quantitative sciences---though notably---we could say significantly---not to large parts of physics and chemistry and geology. We don't claim to understand fully the sciences we survey. But we do understand their unhappy statistical rhetoric. It needs to change.

Contents

Preface

A Significant Problem

In many of the life and human sciences the existence/whether question of the philosophical disciplines has substituted for the size-matters/how-much question of the scientific disciplines. The substitution is causing a loss of jobs, justice, profits, environment, and even life. The substitution we are worrying about here is called “statistical significance”—a qualitative, philosophical rule which has substituted for a quantitative, scientific magnitude.

1. Dieting "Significance" and the Case of Vioxx

Since R. A. Fisher (1890-1962) the sciences that have put statistical significance at their centers have misused it. They have lost interest in estimating and testing for the actual effects of drugs or fertilizers or economic policies. The big problem began when Fisher ignored the size-matters/how-much question central to a statistical test invented by William Sealy Gosset (1876-1937), so-called "Student's *t*.” Fisher substituted for it a qualitative question concerning the “existence” of an effect, by which Fisher meant “low sampling error by an arbitrary standard of variance.” Forgetting after Fisher what is known in statistics as a

"loss function approach,"such as, for example, the minimax strategy, many sciences have fallen into a sizeless stare. They seek sampling precision only. And they end by asserting that sampling precision just *is* oomph, magnitude, practical significance. The minke and sperm whales of Antarctica and the users and makers of Vioxx are some of the recent victims of this bizarre ritual.

2. The Sizeless Stare of Statistical Significance

Crossing frantically a busy street to save your child from certain death is a good gamble. Crossing frantically to get another mustard packet for your hot dog is not. The size of the potential loss if you don't hurry to save your child is larger, most will agree, than the potential loss if you don't get the mustard. But a majority of scientists in economics, medicine, and other statistical fields appear not to grasp the difference. If they have been trained in exclusively Fisherian methods (and nearly all of them have) they look only for a probability of success in the crossing—the existence of a probability of success better than .99 or .95 or .90, and this within the restricted frame of sampling—ignoring in any spiritual or financial currency the value of the prize and the expected cost of pursuing it. In the life and human sciences a majority of scientists look at the world with what we have dubbed "the sizeless stare of statistical significance."

3. What the Sizeless Scientists Say In Defense

The sizeless scientists act as if they believe the *size* of an effect does not matter. In their hearts they do care about size, magnitude, oomph. But strangely they don't measure it.

They substitute "significance" measured in Fisher's way. Then they take the substitution a step further by limiting their concern for error to errors in sampling only. And then they take it a step further still, reducing all errors in sampling to one kind of error—that of excessive skepticism, "Type I error." Their main line of defense for this surprising and unscientific procedure is that, after all, "*statistical* significance," which they have calculated, is "objective." But so too are the digits in the New York City telephone directory, and the spins of a roulette wheel, objective. These are no more relevant to the task of finding out the sizes and properties of soil bacteria or star clusters or investment rates of return than is statistical significance. In short, statistical scientists after Fisher neither test nor estimate, really, truly. They "testimate."

4. Better Practice: Importance vs. "Significance"

The most popular test was invented, we've noted, by Gosset, better known by his penname of "Student," a chemist and brewer at Guinness in Dublin. Gosset didn't think his test was very important to his main goal, which was of course brewing a good-tasting beer at a satisfactory price. The test, Gosset warned right from the beginning, does *not* deal with substantive importance. It does not begin to measure what Gosset called "real error" and "pecuniary advantage," two terms worth reviving in current statistical practice. But Karl Pearson and especially the amazing Ronald Fisher didn't listen. In two great books written and revised during the 1920s and 1930s, Fisher imposed a Rule of Two: if a result departs from an assumed hypothesis by two or more standard deviations of its own sampling

variation, regardless of the size of the prize and the expected cost of going for it, then it is to be called a "significant" scientific finding. If not, not. Fisher told the subjective-phobic sciences that if they wanted to raise their studies "to the rank of sciences" they must employ his Rule. He later urged them to ignore the size-matters/how-much approaches of Gosset, Neyman, Egon Pearson, Wald, Jeffreys, Deming, Shewhart, and Savage. Most statistical scientists have listened to Fisher.

5. A Lot Can Go Wrong with the Use of Significance Tests in Economics

We ourselves in our home field of economics were long enchanted by Fisherian significance and the Rule of Two. But at length we came to wonder why the correlation of prices at home with prices abroad must be "within two standard deviations of 1.0 in the sample" before one could speak about the integration of world markets. And we came to think it strange that the U. S. Department of Labor refused to discuss black teenage unemployment rates of 30-or-40% because with small samples they were, by Fisher's circumscribed definition, "insignificant." After being told repeatedly, if implausibly, that such mistakes in the use of Gosset's test were *not* common in economics, we developed in the 1990s a questionnaire to test for economic as against statistical significance. We applied it to the behavior of our tribe during the 1980s.

6. A Lot Did Go Wrong in the *American Economic Review* during the 1980s

We did not study the scientific writings of amateurs. On the contrary, we studied the *American Economic Review* , a leading journal of economics. With questionnaire in hand

we read every full-length article it published that used a test of statistical significance, January 1980 to December 1989. As we expected, in the 1980s more than 70% of the articles made the significant mistake of R. A. Fisher.

7. Is Economic Practice Improving?

We published our paper in 1996. Some of our colleagues replied, "In the old days [of the 1980s] people made that mistake; but [in the 1990s] we modern sophisticates do not." So in 2004 we published a follow-up study, reading all the articles in the *AER* in the next decade, the 1990s. Sadly, our colleagues were again mistaken. Since the 1980s statistical practice in important respects got worse, not better. About 80% of the articles made the mistaken Fisherian substitution, failing to examine the magnitudes of their results. And less than 10% showed full concern for oomph. In a leading journal of economics, in other words, nine out of ten articles in the 1990s acted as if size doesn't matter for deciding whether a number is big or small, whether an effect is big or small enough to matter. The significance asterisk, the flickering star of *, has become a totem of economic belief.

8. How Big is Big in Economics?

Does globalization hurt the poor, does the minimum wage increase unemployment, does world money cause price inflation, does public welfare undermine self-reliance? Such

scientific questions are always matters of economic significance. *How much* hurt, increase, cause, undermining? Size matters. Oomph is what we seek. But that is not what is found by the statistical methods of modern economics.

9. What the Sizeless Stare Costs, Economically Speaking

Sizeless economic research has produced mistaken findings about purchasing power parity, unemployment programs, monetary policy, rational addiction, and the minimum wage. In truth, it has vitiated most econometric findings since the 1920s and virtually all of them since the Fisher routine got institutionalized in the 1940s. The conclusions of Fisherian studies might occasionally be correct. But only by accident.

10. How Economics Stays That Way: The Textbooks and the Referees

New assistant professors are not to blame. Look rather at the report card on their teachers and editors and referees—notwithstanding cries of anguish from the Savages, Zellners, Grangers, and Leamers of the economics profession. Economists received a quiet warning by F. Y. Edgeworth in 1885—too quiet, it seems—that sampling precision is not the same as oomph. They ignored it and have ignored other warnings, too.

11. The Not-Boring Rise of Significance in Psychology

Did other fields, such as psychology, do the same? Yes. In 1919 Edwin Boring warned his fellow psychologists about confusing so-called statistical with actual significance. Boring

was a famous experimentalist at Harvard. But during his lectures on scientific inference his colleagues appear to have dozed off. Fisher's 5% philosophy was eventually codified by the *Publication Manual of the American Psychological Association*, dictating an erroneous method worldwide to thousands of academic journals in psychology, education, and related sciences, including forensics.

12. Psychometrics Lacks Power

"Power" is a neglected statistical offset to the "first kind of error" of null-hypothesis significance testing. Power assigns a likelihood to the "second kind of error," that of undue gullibility. The leading journals of psychometrics have had their power examined by insiders to the field. The power of most psychological science in the Age of Fisher turns out to have been embarrassingly low or, in more than a few cases, spuriously high—as was found in a 70,000-observation examination of the matter. Like economists the psychologists developed a fetish for testimation, and wandered away from powerful measures of oomph.

13. The Psychology of Psychological Significance Testing

Psychologists and economists have always said that people are "Bayesian learners" or "Neyman-Pearson signal detectors." We learn by doing and by staying alert to the signals. But when psychologists and others propose to test those very hypotheses they use Fisher's

Rule of Two. That is, they erase their own learning and power to detect the signal. They seek a foundation in a Popperian falsificationism long known to be philosophically dubious. What in logic is called the "fallacy of the transposed conditional" has grossly misled psychology and other sizeless sciences. An example is the over-diagnosis of schizophrenia in the United States.

14. Medicine Seeks a Magic Pill

We found that medicine and epidemiology, too, are doing damage with Student's *t* —more in human terms perhaps than are economics and psychology. The scale along which one would measure oomph is very clear in medicine: life or death. Cardiovascular epidemiology, to take one example, combines with gusto the fallacy of the transposed conditional and the sizeless stare of statistical significance. Your mother, with her weak heart, needs to know the oomph of a treatment. Medical testimators aren't saying.

15. Rothman's Revolt

Some medical editors have battled against the 5% philosophy. But even the *New England Journal of Medicine* could not lead medical research back to William Sealy Gosset and the promised land of real science. Neither could the International Committee of Medical Journal Editors, though covering worldwide hundreds of journals. Kenneth Rothman, the founder of *Epidemiology* , brought fantastic progress—to his one journal. His colleagues at other journals looked the other way, as if unaware. Decades ago a sensible few in education, ecology, and sociology initiated a "significance test controversy." But grantors,

journal referees, and tenure committees in the statistical sciences had a strong faith in the idea that probability spaces can judge---the "judgment" merely that $p < .05$ is "better" for variable X than $p < .11$ for variable Y. It's not. It depends on the oomph of X and Y.

16. On Drugs, Disability, and Death

The upshot is that because of Fisher's Standard Error you are being given dangerous medicines, and are being denied the best medicines. The Centers for Disease Control is infected with p-values, in a grant for example to study drug use in Atlanta. Public health has been infected, too. An outbreak of salmonella in South Carolina was studied using significance tests. In consequence a good deal of the outbreak was ignored. In 1995 a Cancer Trialists' Collaborative Group came to a rare consensus on effect size: ten different studies agreed that a certain drug for treating prostate cancer can increase patient survival by 12%. An eleventh study published in the *New England Journal* dismissed the drug. The dismissal was based not on effect size bounded by confidence intervals or what Gosset called "real" error but on a single p-value only, indicating, the Fisherian authors believed,"no clinically meaningful improvement" in survival.

17. Edgeworth's "Significance"

The history of this persistent but mistaken practice is a social study of science. In 1885 an

eccentric and brilliant Oxford don, Francis Ysidro Edgeworth, coined the very term "significance." Edgeworth was prolific in science and philosophy, but was especially interested in watching bees and wasps. In measuring their behavioral differences, though, he focused on the sizes and meanings of the differences. He never depended on *statistical* significance.

18. "Take 3 as Definitely Significant": Pearson's Rule

By contrast, Edgeworth's younger colleague in London, the great and powerful Karl Pearson, used "significance" very heavily indeed. As such things were defined in 1900 Pearson was an advanced thinker—for example, he was an imperialist and a racist, quasi-Fabian, and one of the founding fathers of neo-positivism and of eugenics. Seeking to resolve a tension between passion and science, ethics and rationality, Pearson mistook "significance" for "revelations about the objective world." In 1901 he believed 1.5 to 3 standard deviations were "definitely significant." By 1906, he tried to codify the sizeless stare with a Rule of Three, and tried to teach it to Gosset.

19. Who Sits on the Egg of *Cuculus Canorus*? Not Karl Pearson

Pearson's journal, *Biometrika* (1901-), was for decades a major nest for the significance mistake. An article on the brooding habits of the cuckoo bird, published in the inaugural volume, shows the sizeless stare at its beginnings.

20. Gosset: The Fable of the Bee

Gosset revolutionized statistics in 1908 with two papers published in this same Pearson's journal, "The Probable Error of a Mean" and "The Probable Error of a Correlation Coefficient." Gosset also independently invented Monte Carlo analysis, the theoretical Poisson distribution (unawares), and the economic design of experiments. He conceived in 1926 the ideas of "alternative hypotheses," "power," and "loss," which he gave to Egon Pearson and Jerzy Neyman to complete. Yet most statistical workers know nothing about Gosset. He was exceptionally humble, kindly to other scientists, a good father and husband, altogether a paragon. As suits an amiable worker bee, he planted edible berries, blew a penny whistle, repaired entire, functioning fishing boats with a penknife, and—though a great scientist—was for 38 years a business man brewing beer for Guinness. Gosset always wanted to answer the How Much question. Guinness needed to know. Karl Pearson couldn't understand.

21. Fisher: The Fable of the Wasp

The tragedy in the fable arose from Gosset the bee losing out to R. A. Fisher, the wasp. All agree that Fisher was a genius. Richard Dawkins calls him "the greatest of Darwin's successors." But Fisher was a genius at a certain kind of academic rhetoric and politics as much as at mathematical statistics and genetics. His ascent came at a cost to science—and to Gosset.

Fisher asked Gosset to calculate Gosset's tables of *t* for him, gratis. He then took Gosset's tables, copyrighted them for himself, and in the journal *Metron* and in his *Statistical Methods for Research Workers*, later to be published in thirteen editions and more than a half-dozen languages, he promoted his own circumscribed version of Gosset's test. The new assignment of authorship and the *faux* machinery for science were spread by disciples and by Fisher himself to America and beyond. For decades Harold Hotelling, an important statistician and economist, enthusiastically carried the Fisherian flag. P. C. Mahalanobis, the great Indian scientist, was spellbound.

R. A. Fisher was a necessary condition for the standard error of regressions. No Fisher, no lasting error. But for null hypothesis significance testing to persist in the face of its logical and practical difficulties, something else must be operating. Perhaps it is what Thorstein Veblen called "trained incapacity," to which might be added what Robert Merton called the "bureaucratization of knowledge" and what Friedrich Hayek called the "scientistic prejudice." We suggest that the sizeless sciences need to reform their scientistic bureaucracies.

How, then? Get back to size in science, and to “real error” seriously considered. It is more difficult than Fisherian procedures, and cannot be reduced to mechanical procedures. How big is big is a necessary question in any science, and has no answer independent of the conversation of scientists, as Gosset said. But it has the merit at least of being relevant to science, business, and life. Fisherian procedures are not.

Reader's Guide

Works Cited