A Significant Problem

> *Merely theoretical uncertainty continues to have no meaning. . . . Perhaps as near to it as we can come is in the familiar story of the Oriental potentate who declined to attend a horse race on the ground that it was already well known to him that one horse could run faster than another. His uncertainty as to which of several horses could outspeed the others may be said to have been purely intellectual. But also in the story nothing depended from it; no curiosity was aroused. . . . In other words, he did not care; it made no difference. And it is a strict truism that no one would care about* any *exclusively theoretical uncertainty or certainty. For by definition in being* exclusively *theoretical it is one which makes no difference anywhere.*
>
> JOHN DEWEY 1929, 38–39. EMPHASIS IN ORIGINAL

For the past eighty years it appears that some of the sciences have made a mistake by basing decisions on statistical "significance." Although 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 rule, 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. Our book is not a tract against counting or statistics. On the contrary. In our own scientific work we are quantitative economists and value statistics as a crucial tool.

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, Savage, de Finetti, Good, Lindley, Feynman, Lehmann, DeGroot, Bernardo, Chernoff, Raiffa, Arrow, Blackwell, Friedman, Mosteller, Tukey, Kruskal, Mandelbrot, Wallis, Roberts, Granger, Press, Moore, Berger, Freedman, Rothman, Leamer, and Zellner are quite mistaken—that reducing the scientific problems of testing and measurement and interpretation to one of "statistical significance," as some sciences have done for more than eighty 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 statistical meaning only, 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.

We and our small (if distinguished) group of fellow skeptics say that a finding of "statistical" significance, or the lack of it, statistical *in*significance, is on its own almost 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 the central and standard error of many sciences.

Significance in Science

Statistical significance is the main factual tool of medicine, economics, psychiatry, agronomy, pharmacology, sociology, education, some parts of the biological and earth sciences, and some parts of the academic study of business and history. Astronomers use it to shine a light. Psychologists have developed a fetish for its scientific-sounding rituals. Poll takers and market analysts depend on little else. The tool and its rituals are not much used in the other sciences—atomic physics, say, or cell biology or chemistry or the remaining parts of the life, earth, atmospheric, or historical sciences.

Statistical significance developed first—if comparatively informally—in demography and astronomy. Toward the end of the eighteenth century Laplace began to formalize the notion of significance in the astronomical sciences, though we know from Elizabeth Scott (1953) and others that it functioned in the astronomy of his day as a tiny supplement to scientific reasoning. By the end of the nineteenth century "significance" had been further mathematically refined and extended and began to shed its uncertain light on many fields, both experimental and observational. It became for instance through the works of Galton, Pearson, and Weldon the main instrument of large-sample biometrics.

A brewer of beer, William Sealy Gosset (1876–1937), proved its value in *small* sample situations. He worked at the Guinness Brewery in Dublin, where for most of his working life he was the head experimental brewer. He saw in 1905 the need for a small-sample test because he was testing varieties of hops and barley in field samples with *N* as small as four. Gosset, who is hardly remembered nowadays, quietly invented many of the tools of modern applied statistics, including Monte Carlo analysis, the balanced design of experiments, and, especially, Student's *t,* which is the foundation of small-sample theory and the most commonly used test of statistical significance in the sciences.[1] Gosset's "The Probable Error of a Mean" was published in 1908 under the pseudonym "Student." Yet he had been thinking about his need for a small sample test since at least 1905, the year he told Karl Pearson about it. But the value Gosset intended with his test, he said without deviation from 1905 until his death in 1937, was its ability to sharpen statements of *substantive* or *economic* significance. Gosset's immediate goal, of course, was to brew the best tasting stout at a satisfying price. His test of statistical significance could, he knew, contribute only a little to those substantive aesthetic and economic goals. Experiments in the selection and cultivation of barley and hops, and in quantitative simulations, technologies of malting, water quality, yeast chemistry, storage temperature, cask type, and many other Guinness variables—from an unusually generous if paternalistic wages and benefits scheme to the daily taste test—would contribute far more.[2] World War I had been under way for more than a year when Gosset—who wanted to serve in the war but was rejected because of nearsightedness—wrote to his elderly friend, the great Karl Pearson: "My own war work is obviously to brew Guinness stout in such a way as to waste as little labor and material as possible, and I am hoping to help to do something fairly creditable in that way."[3] It seems he did.

Counting Matters, and Inference, Too

Every science uses counting and should. Counting is central to a real science, and applied statistics is sophisticated counting. The big scientific question is, "How much?" To answer the how-much question you will often need statistical methods. If your clinical-diagnostic problem can best be answered with "analysis of variance"—though we advise you to think twice—then you had better not be terrified of its imposing columns and nonlinear off-diagonals. If your business problem is a sampling one, and requires a high degree of confidence in the truth of the result, you had better examine the entire "power function." If your physical problem leads naturally to the Poisson distribution you had better be fluent in that bit of statistical theory. If you want to say how much abortion reduces crime rates, "other things equal," you had better know how to run multiple regressions that can isolate the effect of abortion from those other things. Good.

Statistical significance is a subset of such statistical methods. Formally speaking, statistical significance is a subset of induction, either "inductive behavior" (the question of how much) or "inductive inference" (the question of whether), depending on one's philosophical school of thought.[4] But statistical "significance," once a tiny part of statistics, has metastasized. You can spot statistical significance in the sciences that use it by noting the presence of an F or p or t or R^2—an asterisk superscripted on a result or a parenthetical number placed underneath it, usually with the word *significance* or *standard error* in attendance. Scientists use statistical significance to "test" a hypothesis—for example, the hypothesis that comets come from outside the solar system or the hypothesis that social welfare programs diminish the pace of economic growth.[5] So statistical significance is also a subset of testing.

Testing, too, is used by all the sciences, and of course should be. That is, claims should be estimated and tested. Scientific assertions should be confronted quantitatively with the world as it is or else the assertion is a philosophical or mathematical one, meritorious no doubt in its own terms but not scientific. To demonstrate scientifically and statistically that Chicago is the City of the Big Shoulders you would need to show by how much Boston or London lack such shoulders.

The problem we are highlighting is that the so-called test of statistical significance does not in fact answer a quantitative, scientific question. Statistical significance is not a *scientific* test. It is a philosophical, qualitative

test. It does not ask how much. It asks "whether." Existence, the question of whether, is interesting. But it is not scientific.

The question of whether is studied systematically in departments of philosophy or mathematics or theology. We have spent time in those departments and have a high opinion of them. But their enterprises are not scientific. Instead of asking scientifically how big *are* the shoulders of Chicago, the philosophical disciplines ask *whether* big shoulders *exist*. Does there exist an obligation to believe a synthetic proposition? Yes or no. Does there exist a good and omnipotent God? Yes or no. Does there exist an even number not the sum of two primes? Yes or no. Does there exist a significant relationship between economic growth and belief in hell? Pray tell.

The problem we are pointing out here—we note again that it is well known by sophisticated students of the matter and is extremely elementary—is that by using Fisherian methods some of the putatively quantitative sciences have slipped into asking qualitatively whether there *exists* an effect of drug prices on addiction or whether there *exists* an effect of Vioxx on heart attacks or whether there *exists* an effect of Catholicism on national economic backwardness. Yes or no, they say, and then they stop. They have ceased asking the scientific question "How much is the effect?" And they have therefore ceased being interested in the pragmatic questions that follow: "What Difference Does the Effect Make?" and "Who Cares?" They have become, as we put it, "sizeless."

But the Point of Counting and Inference Is to Find a Size

Real science depends on size, on magnitude. Scientific departments of physics, economics, engineering, history, medicine, and so forth intend to study actualities and realistic possibilities quantitatively. We admire their quantitative intentions, in many fields first imagined in part by statistical sophisticates such as Francis Galton and Karl Pearson. Victor Hilts, the historian of science, wrote in 1973, "I think it might even be fair to say that the introduction of statistical techniques in the social sciences represents one of the two most important methodological innovations in [all of] nineteenth-century science"(Hilts 1973, 207). We agree.

But after Galton and Pearson, and especially after Ronald Fisher, the statistical sciences have slipped into asking a philosophical and qualitative question about existence instead. The scientific question is how much this

particular bridge, or a bridge of this particular kind, can tolerate thus-and-such forces of stress. There may "exist" a stable bridge. But unless the magnitudes and limits of stability can be given quantitatively in the world we actually inhabit the knowledge of whether it exists is unhelpful. No astronomer is interested in the question of whether there is some effect of the rest of the galaxy's gravitation on the Oort cloud. No scientific brewer of Guinness will ask whether bitterness "exists"—as a careful student of hops chemistry, and a profit center, he is forced to ask how much. The question of whether has, as John Dewey observed, "no [scientific] meaning," no big bang. Being "exclusively theoretical" no curiosity is aroused by it because it makes "no difference" anywhere. Not even in philosophy, pragmatically considered, the great philosopher said.

Like an engineer the astronomer is interested in *how much* the effect on the Oort cloud is, for the generation of comets, say. Instead of *nonexistent* or *not statistically significant* she uses a word seldom heard in the ontological and metaphysical departments of philosophy and mathematics and theology: *negligible*. In a department of mathematics it would be viewed as irrelevant, even vulgar, to note that the number of even numbers not expressible as the sum of two primes is entirely negligible—as in fact it appears to be. No such even number has yet been found, up to gigantic numbers. But that, the mathematician will complain with a sneer, is a mere calculation, satisfactory for mere engineers and physicists but not for a "real" mathematician. The real mathematician, since the Greeks first invented such a character, has craved certitude. But certitude is not interesting to the astronomer. She seeks magnitude and effect size. She is seldom tempted to substitute "testing for low magnitude" or "estimation of effect size" for "significance testing for a low probability." In a science, size matters.

To confront her assertions with the world the astronomer or engineer uses simulation methods, for which often enough no unique analytic solution is guaranteed to "exist." She is not against analytic certitude, whatever that might mean. Her quantitative methods—estimating magnitudes—puzzle her colleagues in mathematics and, more to the point here, puzzle, too, her colleagues in econometrics or statistical medicine who have spent too much time in the exist/not-exist world of esteemed departments of mathematics, philosophy, or theology. The astronomical scientist wants to meet scientific, not metaphysical, standards. She seeks salience, adequacy, nonnegligibility, real error, an oomph that measures the practical difference something makes. Such scientific standards of per-

suasion have little to do with axiom or consistency or existence. Axiom, consistency, and existence are the values of the nonscientific departments, admirable in themselves, we repeat earnestly. But they are not the values of a quantitative science.

The substitution of existence for magnitude has been a grave mistake. That's one way of stating the main point of our book. Any quantitative science answers how much, or should. Many do, like geophysics and chemistry and a good deal of history. How much energy did the crashing of the Indian subcontinent contribute to the raising of the Himalayas? How much did foreign trade contribute to the British industrial revolution? How much genetic material is transmitted to the next generation? But medicine, economics, and some other sciences have stopped asking how much, especially in their academic, as against their applied, work. Or, to be more exact, they believe they *are* interested in the quantitative questions of what this or that number really is in the world. But their way of deciding what that number really is—statistical significance or insignificance—and what difference it makes—doesn't give them the correct answer.

Statistical significance sounds scientific. After all, it speaks in technical terms about the experimental and observational quantities of science. And it shows up in, for example, the technical notes to a bottle of prescription pills. "Zyprexa was significantly better than placebo ($P = 0.05$) on initial assessment," says the guardian of the quantitative bottom line behind an antidepressant pill, "with the Cox model but not when compared with placebo by means of a Hochberg adjustment for multiple comparisons" (Osterweil 2006, 2). How very significant.

But in truth statistical significance is a philosophy of mere existence. And even by philosophical standards—leaving aside the scientific standard—it is a poor one. It concerns itself only with one kind of probability of a (allegedly) randomly sampled event—the so-called exact p-value or Student's t—and not with the other kinds of sampling probability, such as the "power of the test," which controls for what Egon Pearson called in 1928 the "second type of error." And statistical significance is not concerned with any of the long list of *nonsampling* sources of error, such as confounding effects, as one finds in medicine and epidemiology; specification error, as one finds in economics and the other human sciences; "non-linear fertility slopes" (Student 1938, 374), as one finds in agronomic experiments; the "bias of the auspices" (Deming 1950, 43), as one finds in government, industrial, and ethnographic studies; measurement

error, as one finds in psychology and pharmacology research and everywhere else; or experimental error and sample selection bias, as again one finds in all of the sciences.[6] Significance is a strikingly partial philosophical account of the existence of error, and Fisher's version of it—sans power—is lacking even a probabilistic means of assessing its own magnitude. It fails, as the great geo- and astrophysicist Harold Jeffreys reminded Fisher, to "give us ground for believing the laws that we do believe or else say definitely that our inferences are fallacious."[7]

But the biggest problem is the more elementary one we shall explore here. It is: *Fit is not the same thing as importance*. So-called tests of fit, such as Student's *t* or tests of R^2, do not by themselves solve the all important matter of *Gosset* significance—which is Size Matters + Who Cares?—the "minimax strategy" or "loss function" of every inquiry.[8] Scientists and their customers wish to have relevance not amateur philosophy. They wish for standards that will help them, say, minimize the maximum loss of jobs, income, profit, health, or freedom in following this or that hypothesis as if true.

The general validity of the loss function way of thinking is, incidentally, not sensitive to the degree of risk aversion felt by an individual investigator or his advisees. So "personal taste" or "equal distribution of ignorance" are not excuses justifying gross indulgence of the mechanical instrument. Loss is in the behavioralistic tradition of statistics the perceived value of a sacrifice, giving up that to do this, and may be positive or negative. So even the gambler who sees "nothing but blue sky"—a real risk lover—will distinguish maximum losses from minimum, big wins from small. Adjusting the levels of Type I and Type II error is, statisticians agree, necessary for handling differential attitudes toward risk. The problem is that today's statistical experts do not estimate or consider the loss function or Type II error at all. In his last year of life, the great statistician and economist Leonard Savage asked, "When is one [statistical] expert, real or synthetic, to be preferred to another?" He replied, "Employ, until you have further experience, that expert whose past opinions, applied to your affairs, would have yielded you the highest average income" (1971b, 145–46). Substitute "highest average income"—or rather add to it—other concerns, such as "highest average quality" or "highest rate of patient survival" or "lowest number of heart attacks" or "highest average rate of minority student graduations," or even "highest scientific consensus" and you have what we are claiming here.

In any case, without a loss function a test of statistical significance is

meaningless, no better than a table of random numbers. Pretending to afford a view from everywhere, statistical significance is in fact a view from nowhere. In its desire to maximize precision in one kind of sampling error it turns away from the human purposes and problems that motivated the research in the first place. Fisher-significance is by itself about precisely nothing.

Statistical significance has translated every quantitative question about hypotheses into a philosophical and qualitative measure of probability about the data assuming the truth of a singular hypothesis. It has collapsed the scientific world into a Borel space, p (0, 1.0)—a procedure, by the way, that the mathematical statistician Émile Borel (1871–1956) himself emphatically rejected. Borel, though a master of abstract imagination, was deeply interested in the substantive side of testing and in Paris in the 1920s helped convert a young Jerzy Neyman to a life of substantive significance.[9]

Savage noted in *The Foundations of Statistics* (1954) a part of the problem we are highlighting: "Many [scientists following in the footsteps of Karl Pearson and R. A. Fisher]," he wrote, "have thought it natural to extend logic by setting up criteria for the extent to which one proposition tends to imply, or provide evidence for, another. . . . It seems to me obvious, however, that what is ultimately wanted is criteria for deciding among possible courses of action."[10] Yet, in imitation of Fisher, today's significance testers do not think about "possible courses of action." Statistical significance is, as Savage says, "at best a roundabout method of attack."[11] That is to put it charitably.

To cease measuring oomph and its relevant sampling and nonsampling error is to wander off into probability spaces, forgetting—commonly forever—that your interest began in a space of economic or medical or psychological or pharmacological significance. In applied work at, for example, the Pfizer Corporation or the Centers for Disease Control or the Federal Reserve Board of Governors, one would expect the scientists involved to be serious about substantive oomph in medical or economic or pharmacological matters. But p and t and F and other measures of "significance" fill the air.

The sociological pressure to assent to the ritual is great. In 2002 we gave together a talk at the Georgia Institute of Technology, where Ziliak was teaching, on the significance mistake in economics. Three researchers from the nearby Centers for Disease Control (CDC) attended. They agreed with us about "the cult of p," as they put it. But they feared that

their mere presence at a lecture against Fisher's "significance" would put their jobs at risk and made us promise not to reveal their names. The official rhetoric at the CDC is: "Second-hand smoke is killing thousands annually. But is it *statistically* significant?" The CDC, which is unquestionably one of the world's most important sites for the scientific study and control of disease, imposes Fisherian orthodoxy on its scientists. That is the power that the cult of statistical significance has.

Sizelessness is not what most of the Fisherians believe they are getting. The sizeless scientists have adopted a method of deciding which numbers are significant that has little to do with humanly significant numbers. The scientists are counting, to be sure: "3.14159***," they proudly report, or simply "***." But, as the probabilist Bruno de Finetti said, the sizeless scientists are acting as though "addition requires different operations if concerned with pure numbers or amounts of money" (De Finetti 1971, 486, quoted in Savage 1971a).

Substituting "significance" for scientific how much would imply that the value of a lottery ticket is *the chance itself,* the chance 1 in 38,000, say, or 1 in 100,000,000. It supposes that the only source of value in the lottery is sampling variability. It sets aside as irrelevant—simply ignores—the value of the expected prize, the millions that success in the lottery could in fact yield. Setting aside both old and new criticisms of expected utility theory, a prize of $3.56 is very different, other things equal, from a prize of $356,000,000.[12] No matter. Statistical significance, startlingly, ignores the difference.

Imagine that you and your infant child are standing on a sidewalk near a busy street. You have just purchased a hot dog from a street vendor, and have already safely crossed the street. You suddenly realize that you've forgotten the mustard. Prize Number One: if you and your child scurry across the busy street, dodging moving trucks and cars, there is some probability—say, 0.95—you'll both safely return with the mustard in hand.

Now imagine that you've gotten the hot dog and mustard across the street all right—but this time you've forgotten your infant child. You watch in horror as she tries to cross the street by herself. Prize Number Two: if you scurry across the street, dodging vehicles as before, there is some probability—say, the same 0.95—you'll return with your child unharmed.

Two prizes—the mustard and your child—identical probability. Statistical significance ignores the difference. It ignores, in the words of Savage, "criteria for deciding among possible courses of action." Since both decisions are equal in probability of "success," that is, equal in *statistical*

significance, the sizeless scientists assign equal value to the mustard and the child. Both the CHILD variable and the MUSTARD variable are significant at $p = .05$. Therefore, the sizeless scientist in effect declares, "They are equally important reasons for crossing the street."

Imagine a crime suspect interrogated repeatedly by investigators. Significance testers say that if ninety-five times or more out of a hundred a suspect admits he is guilty of the crime, the correct decision is "guilty," hang him high, regardless of the size or nature of his alleged crime, ax murder or double parking, and the ethical and pecuniary cost to society of disposing of him. But if his testimony is the same only ninety-four times of out a hundred, or eighty-one times out of a hundred, let the suspect go unmolested. He's innocent, again despite the size, nature, or cost of the alleged crime. If the p doesn't fit—the sizeless scientists say, following the dictates of R. A. Fisher—then you must acquit.

William James believed that "we have the right to believe any hypothesis that is live enough to tempt our will" (1896, 107) but added—in a philosophical pragmatism friendly to what Neyman and Savage called "inductive behavior"—what really tempts is what a belief "leads to" (98). What, a scientist should ask, are the social or personal human purposes activated by the belief? What does the belief lead to? A conclusion such as "the subjects of my inquiry are indifferent between a mustard packet and a child" will rarely lead one to the correct side of the street. A sizeless scientific finding, measuring nothing relevant to human decision making, leads nowhere—it is therefore, in strictest pragmatist terms, as the philosophers Quine and Ullian put it, "unbelievable."[13]

The usual procedure does not ask the question "How big is big?" about its numbers . It does not ask whether the variable is for human or other purposes substantively significant. It asks instead, "In the data we happen to have, is the estimated variable (or the full model, if that's what is being tested) more than two or three standard deviations of its own sampling variation away from the null hypothesis?" It's not the main question of science. But it is the only question that a sizeless scientist bothers with. "*X* has at the .05 level a significant effect on *Y,*" he says. "Therefore *X* is important for explaining *Y*." The circumlocution is made regardless of how much a unit increase in *X* affects the levels or qualities of *Y* or how many other variables are involved or how much *X* varies or what difference *X* makes in the course of *Y* when *Z* is added to the experiment. Dewey wrote that "it is a strict truism that [no scientist] would care about *any* exclusively theoretical uncertainty or certainty. For by definition in

being *exclusively* theoretical it is one which makes no difference anywhere" (1929, 38–39, emphasis in original). Yes.

When the Fisherian significance tester wishes to use a set of data to distinguish between two different hypotheses, such as $\beta = 1$ (the null hypothesis, say) and $\beta > 1$ (the one-sided alternative), he asks *whether* the variate is "statistically significantly different from the null" (which null hypothesis he assumes provisionally to be true). Exclusively he asks about the theoretical certainty or lack of it. He does not ask the relevant scientific question—how much oomph, how much pragmatic effect, what difference does the variate or model lead to relative to some different magnitude of scientific importance? Does a β of 1.20 lead to other economic or medical recommendations? Does a β of 1.10 tempt our will? What are the available courses of action and how do you know?

In a regression context, an estimated $\beta_{hat} = .01$ (to take another magnitude) is said by such a scientist to be *different from* a theoretical null hypothesis β_{null}, equal to, let's say, exactly zero, if with a large enough sample and a small enough variation in the sample at hand the variate attached to β_{hat} is "statistically significantly different from the null hypothesis of [exactly] zero." That the independent variate in question is "amount of insulin dose" and the dependent variate "length of patient survival" does not affect procedures. The Fisherian procedure claims to test the significance of numbers "in their own terms," objectively, without regard for human purposes. (*In*significance, we should add, a failure to achieve the .05 or .01 cutoff, is the other side. "The coefficient on DIALYSIS [or MONEY or TAXES or DEATH—the list of candidates is endless] is statistically *in*significant but of the right sign," as a great many authors have written in slavish imitation of each other. In our home field of economics we call such practice "sign econometrics." It is rampant.) But the equation and substitution of statistical significance and scientific relevance are proposed relative to no scientific or ethical values. Literally, we repeat, none.

A book could be excellently copyedited, precise in every detail, the publisher and the author having spent months and months, thousands of man hours, making sure that every number in the book is precise to eight significant digits, every jot and tittle of every word just so, a very Torah scroll of precision. Yet the book could be substantively worthless, a book, say, consisting of Fisherian significance tests on numbers gathered with no scientific question in mind from telephone directories and MySpace sites in Holland, Greece, and Tanzania. On the other hand, an important book,

such as Fisher's *Design of Experiments* (1935), could be imprecise in very many details, with notable misspellings, say, and errors in the statistical tables, illogicalities in the mathematics, or, worse (as we claim was in fact the case for *The Design of Experiments*), it could have vitiating errors in its scientific rhetoric, yet it could nonetheless be important in the intellectual history of the twentieth century and well worth publishing. No sensible person—or publisher—would confuse precision in production with importance in intellectual life. Yet that is what users and buyers of tests of significance do.

Precision in the matter of random sampling is nice, to be sure, something to be desired. Sometimes having "precision" narrowed to "precision in the sense of an arbitrarily low standard error from an alleged sample of an imaginary repeated experiment in view of an arbitrary null hypothesis of exactly zero" is scientifically useful. But rarely. The problem is the opportunity cost of specializing in it excessively, as Fisherian procedures urge one to do. A publisher would never say to herself, "This book is an idiotic compilation of significance tests on telephone directories and MySpace sites that leads nowhere and tempts no will. But after all it is beautifully and precisely copyedited, with every number and spelling and citation checked fifty times. Therefore I have done my job." But that, alas, is what the sizeless sciences say to themselves, especially since the arrival of the desktop computer with its ability to invert big matrices at the punch of a key, "checking" on sampling variability effortlessly and on a gigantic scale. By Moore's law electronic computation of statistical significance has cheapened to near zero. By economic law the scientific value of the computation has come to be equal at the margin to its private cost. "Decision" has become socialized and bureaucratized—heedless of the social margins. The sciences of medicine and economics and the others have developed machinery heedless of scientific substance.

The substitute question is supposed to tell "whether" an effect "exists." It does not. If beneath your β's you can insert $p < .05$ or $t > 2.0$, then the scientific job is supposed to be finished regardless of effect size and its relevance. It is not.

The Sizeless Scientists Have Missed the Point

The problem of significance is old. The substitution of significance for relevance is more than a century old in some sciences, such as parts of biology—anthropometry, for example. And since Ronald Aylmer Fisher made

it canonical in the 1920s and 1930s the standard error has spread and spread. The idea of significance as applied to matters of random sampling existed, we have noted, in the early eighteenth century, originating, it would appear, with John Arbuthnot and Daniel Bernoulli. The very word *significance* seems to have first appeared in an article by F. Y. Edgeworth, in 1885. The basic logic is ancient and has, we've said, its uses. Cicero's Quintus character in *De Divinatione* uses it to argue for the existence of the gods: "'Mere accidents,' you say. . . . Can anything be an 'accident' which bears on it every mark of truth? Four dice are cast and a Venus throw [four different numbers] results—that is chance; but do you think it would be chance, too, if in one hundred casts you made one hundred Venus throws?" (I, 23). He's got a point.

But significance was a minor part of any science until Francis Galton (1822–1911) and especially Karl Pearson (1857–1936) fitted *Biometrika* with chi-square, regressions, and other curves. After Fisher published *Statistical Methods for Research Workers* in 1925, among the most widely read professional texts on statistics of the twentieth century, statistical significance became the central empirical question, commonly the only empirical question—first in biometry and agronomy, then in genetics, psychology, economics, anthropometry, sociology, medicine, law, and many other fields. In psychometrics, statistical significance was by the late 1920s—regardless of effect size—judged a necessary and sufficient condition for the demonstration of a scientific result. By the 1970s Fisher's absolute criterion of significance was flourishing in all the fields we highlight here and was sinking its roots into a few more that we have only glanced at, such as law.

Fisher's procedure of statistical significance often has other difficulties, on which a great deal of philosophical and mathematical and sociological ingenuity has been spent. Is the sample proper? Is the whole population in the so-called sample? Is the "sample" one of convenience, and therefore biased in preselection? Has the investigator herself selected for use only the significant results ($p < .05$) out of many experiments attempted? Is the significance level corrected for sample size? Is the assumed sampling distribution the correct one? What, after all, is the correct measure of probability? Is probability about *belief*—in De Morgan's terms, a "law of thought"—or is it mainly about frequencies in the long run—a "law of things," as John Venn believed?[14] What is the relation between "personal" probability and groupwise calculable risk?[15] What about heteroskedacity? What about truncation error? What about spec-

ification error? And so on. You may consult many thousands of excellent philosophical and statistical articles published each year that examine these absorbing difficulties.

But we are making a more elementary point. Statistical significance, we are saying, is never the end of an argument. Indeed, speaking of stepwise procedures, it's always a false start. Statistical significance is neither necessary nor sufficient for a scientific result. Many who have seriously examined the issue, from Edgeworth to Kruskal, agree. Statistical significance offers merely a certain kind of theological proof: ipse dixit—a *t*-statistic is supposed to "speak for itself" because the probability of a certain restricted kind of error is low, a sampling error of excessive skepticism. Most scientists seem to believe this. In grant applications as much as journal correspondence, they verbally repeat the belief as if it were the very paternoster of science. We want to persuade them to go back to a truly scientific ritual, asking How Much.

Science Needs a Loss Function

The doctor who cannot distinguish statistical significance from substantive significance, an *F*-statistic from a heart attack, is like an economist who ignores opportunity cost—what statistical theorists call the loss function. The doctors of "significance" in medicine and economy are merely "deciding what to say rather than what to do" (Savage 1954, 159). In the 1950s Ronald Fisher published an article and a book that intended to rid *decision* from the vocabulary of working statisticians (1955, 1956). He was annoyed by the rising authority in highbrow circles of those he called "the Neymanites."

But every inference drawn from a test of statistical significance is a "decision" involving substantive loss and, further, not merely one narrow sort of loss under conditions of random sampling. Every decision involves cost and benefit, needs and wants, choices and courses, a minimax problem (if that is your loss function) as general or particular as the problem-situation warrants. Accepting or rejecting a test of significance without considering the potential losses from the available courses of action is buying a pig in a poke. It is not ethically or economically defensible.

We want you to be dissatisfied with a 5 percent "verbalistic" philosophy (Savage 1954, 159). The solution is not to seek a 1 percent or a 12 percent philosophy. No predetermined rule of one or three or whatever for α-level sampling error will do, as Gosset said repeatedly to Fisher, Karl

Pearson, Egon Pearson, Edwin S. Beaven, and others. Significance rules in isolation are useless.

All the signals point you toward seeking instead, speaking in regression terms, a 100 percent β-philosophy. To reclaim the quantitative side of your science you will need to make β-decisions: you will have to make β-decisions about differences expressed in terms of effect size and, if you are seriously worried about sampling as one source of error, you will have to make β-decisions about your power to reject the null relative to substantive alternatives. You will have to think about your coefficients in a currency of How Much in the world as it is, or could be, and persuade a community of scientists. Instead of deploying a mechanical rule about one kind of sampling error you will have to establish a reservation price of β-coefficients, a minimum effect size of substantive significance, in the relevant range of power, for your particular area of research, acknowledging all the sources of error. You will not confuse power with substance, nor mere sampling error with "real" or "actual" error (Student 1927, 1938). You will instead dwell on the substantive meaning of your estimates in the range of real error. That is, you will *actually* repeat experiments, as Gosset did, not pretend to, as Fisher and his many followers have, so that your sampling distribution is based on something besides an imagined infinitely repeated flipping of a fair coin. You will employ minimax or some other loss function to consider the ramifications of possible courses of action or interpretation. You will give an economic interpretation to the logic of uncertainty. In the style of Gosset you will supply "*real* error bars" around your best estimates, showing that sampling-based confidence intervals are only one element—perhaps a quite small element—of the discussable error.[16] You will, in other words, draw a dividing line of believable effect size at which some phenomenon should be considered scientifically or humanly important. You will devote your energies to examining the substantive deviations from this minimum oomph.

A tall order? Yes, but it has an honored name. It is called "science." Variable by variable, model by model, it is a difficult change to make, from the nonscience of statistical "significance" to an actual science of oomph. It moves away from the metaphysics of "existence" characteristic of Greek-derived mathematics and philosophy to the calculable magnitudes characteristic of modern science since the seventeenth century. It's hard to do, unlike calculating *t*-statistics, which is a simpleton's parlor game. But actual science at the frontier is supposed to be difficult. If it wasn't, you wouldn't be at the frontier.

The other problem we are highlighting is what is known as the *fallacy of the transposed conditional.* None of the oomph-lacking tests of Student's *t,* as its inventor told Fisher and Egon Pearson in 1926, are logically speaking tests of hypotheses at all. Fisher is the reason. He erased the Bayesian odds from Gosset's original test of hypotheses, ignored Gosset's and Neyman's and Pearson's insights about the power of the test, and instead calculated the likelihood of observing the given data, assuming a single null hypothesis is true. In some instances of science Fisher's inversion may pose no particular problem. It depends on the question one is asking. But in daily use of Fisher's methods the logic is turned on its head: the sizeless scientists claim to observe the likelihood of a null hypothesis, assuming the data they happen to have in hand are true, the exact reverse of what Fisher's method produces. "If *H,* then *O*" is supposed to affirm "If *O,* then *H.*" It doesn't.

If a person is hanged, he will probably die. Therefore, say the sizeless scientists on coming upon a corpse in the street, "He was probably hanged." Something is amiss. The probability of being dead, given that you were hanged, is much higher—much more *statistically* significant, as irrelevant as such a proposition is for finding out exactly why a person is dead—than the probability that you were hanged given that you are dead. This is the fallacy of the transposed conditional. A high likelihood of the sample, supposing the hypothesis is true, is supposed to imply a high probability of the truth of the hypothesis in light of the sample. No. Bringing back Bayes's rule, as sophisticates such as Lindley and Zellner and Good and others have done, is probably a good idea (Bernando 2006). But on our two main points a Bayesian revolution is not necessary, merely efficient. What *is* necessary is to clearly distinguish Gossetian *hypothesis* testing from Fisherian *significance* testing.

In other words, since the 1920s many economic and medical and psychological and forensic scientists have calculated the wrong probability. To take one of by now literally millions of examples, the fallacy of the transposed conditional has in psychometrics led to a gross overestimate of the number of adults afflicted by schizophrenia (Cohen 1994, 999–1000). You can well imagine the need for and relevance of the size-matters/how-much question in all of the sciences.

After Fisher, then, the sizeless sciences neither test nor estimate. They practice a third method of science not easily recognized by a Fisher-only education. The third method—which is a marriage of the sizeless stare of statistical significance to the fallacy of the transposed conditional—we call *testimation,* the ruin of empirical research.

Fisher's testimation has arrived at high places, such as the Supreme Court of the United States. In *Castenda v. Partida,* 430 U.S. 482 (1977), concerning jury discrimination, the court held that "as a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations [that is, if $t > 2.0$ or 3.0 or $p < .05$ or .01], then the hypothesis would be suspect to a social scientist, 430 U.S. at 496 n.17."[17] This is mistaken. By God's grace the estimate of jury discrimination in question may be good or bad. But its sheer statistical significance is no evidence one way or the other. If the variable doesn't fit / You may not have to acquit. It depends on the oomph, the expected loss of sticking to the null. A suspect and his jury deserve to get from the expert a clear statement about the warrants of maintaining one hypothesis over another given the truth of the observed data. Today they get instead a statement about the warrants of maintaining the observed data, assuming the truth of a maintained hypothesis (the null hypothesis of, say, "innocent"). The law and all such work—that is to say, most of the statistical work in economics, psychology, medicine, and the rest since the 1920s and especially since the 1980s and the coming of personal computers—has to be done over again.

But the Standard Error Is Tempting: Gosset Knew

Gosset himself never believed "significance" was 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 an apprentice brewer in the Experimental Division (1899–1906), later head experimental brewer (1907–35), and for the rest of his short life head brewer of Guinness in both Dublin and the newly established brewery in London, Park Royal (1935–37).[18] Gosset's field experiments with barley varieties in Ireland and England yielded small samples. Agricultural experiments are expensive, he knew firsthand, so large sample sizes, on which statistical theory since the eighteenth century had been based, were not profitably relevant. How then could he distinguish the mean difference between, say, two barley yields with $N = 4$? Even with small samples the evidence for or against a null hypothesis of no difference was, Gosset told Karl Pearson in an important 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 in mathematics [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, ca. April 1905, in E. Pearson 1939, 215–16; italics supplied)

Pearson didn't understand the advice Gosset was requesting and certainly never realized that Gosset was the one who was giving the advice. The great man of large samples never did grasp Gosset's point—though wisely he agreed to publish "Student's" papers.

Gosset seems never to have tired of teaching it. Twenty-one years after his letter of 1905 he responded to a query by Egon Pearson (1895–80), the eldest son of the great Karl, who, unlike Pearson *père,* definitely did grasp the point. In his response Gosset 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" [measured in beer bitterness or lives or jobs or pounds sterling] relative to some "alternative hypothesis."[19] 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 relevant.

Gosset's letters to the two Pearsons, and his twenty-one published articles, we have noted, are essentially unknown to users of statistics and especially to economists. Yet Gosset was proposing, and using in his own work at Guinness, a characteristically economic way of looking at the acquisition of knowledge. He focused on the opportunity cost, the value of the sacrifice incurred by choosing one of the competing hypotheses. It became the way of Neyman-Pearson and Wald and Savage, though crushed in practice by Fisher's forceful, antieconomic campaign. Fisher, by contrast, looked for some absolute qualitative essence in the aristocratic style of philosophy or theology or even of economics itself in the age before the idea of opportunity cost was made clear.[20] After the Gosset letters of May 1926 Egon Pearson and Jerzy Neyman began a series of famous theoretical papers establishing beyond cavil that Fisher was wrong and Gosset right. To no avail.

"*Pecuniary* advantage," Gosset's invention, is not the unique currency of a regression coefficient or the difference between two means. Gosset, a man of sense and compassion, well understood this—though even today in cancer epidemiology the merely financial expense has a partial claim, considering the alternative employment of the money in saving other lives. Gosset himself conducted original studies on genetics, yeast, barley, hops, the water of the Thames, and, years later, as we have noted, the nutritional advantage to children of drinking raw milk (Student 1931a). He focused always on the substantive meaning—the *chemical* and *biological* significance of his coefficients. If you yourself deal in medicine or psychiatry or experimental psychology, Gosset and we would recommend that you focus on *clinical* significance. If you deal in complete life forms, *environmental* or *ecological* significance. If you deal in autopsies or crime or drugs, *forensic* or *psychopharmacological* significance. And so forth. In short, Gosset's rule is: in any science, attend to oomph. An arbitrary and Fisherian notion of "statistical" significance should never occupy the center of scientific judgment.

A great yet essentially unknown scientist, Gosset quietly invented, among other things, the definition of economic significance, the statistical "design of experiments," the table of *t,* the *t*-test, and even the ideas of "alternative hypotheses," "power," and "loss" (Ziliak 2008a). He did not consider economic calculation or power or loss as mere add-ons or optional accoutrements in case "you have time" or "are curious" after grinding out significant *t*'s and high R^2. Gosset shrugged at a merely statistical significance found in the single sample on offer. Late in life he wrote a letter to Egon, who had recently succeeded his father as editor of *Biometrika*.[21] Gosset was working on experiments with another old friend, Edwin S. Beaven (1857–1941), a pioneer in agricultural experiments and the world's leading authority on barley and malt.[22] Gosset wished to publish in *Biometrika* the results of their experiments together with his own latest thinking about the role of "significance" in the design of experiments. "The important thing in such," Gosset wrote, "is to have a low real error, not to have a 'significant' result at a particular station [as Fisher sought]. *The latter*"—that is, a merely statistical significance defined in Fisher's way, he told the new editor of *Biometrika*—"*seems to me to be nearly valueless in itself.*"[23]

Gosset was prophetic against the mechanization of statistical instruments, too, including even calculating machines. He computed the table of *t* with a mechanical calculator, "Baby Triumphator," motored by a

turn-crank and his own strong arm. But he had used electric machines, too, at the brewery and later at Fisher's office, and felt the intellectual difference. The same dangerous ease of calculation has brought statistical significance to a peak in our own Early Computer Age. In Gosset's 1905 report on "The Pearson Co-efficient of Correlation," he warned his fellow brewers that "the better the instrument the greater the danger of using it unintelligently. . . . Statistical examination in each case may help much, but no statistical methods will ever replace thought as a way of avoiding pitfalls."[24] Statistical instruments, such as the *t*-test, the correlation coefficient, and Intel Inside/Celeron, will not replace thought. Precisely. Some years later the poet and Latin textual critic A. E. Housman complained about the replacing of thought with thoughtlessly mechanical rules (for example, "Honor the existing texts even if they yield nonsense") in an article entitled, with heavy sarcasm, "The Application of Thought to Textual Criticism" (Housman 1922 [1961]; see McCloskey 1985 [1998], 72–73). Gosset and we would like to see the application of thought to statistical methods.

But Gosset was not as forceful as Housman or Fisher. He had, a friend of his schooldays said, "an immovable foundation of niceness." He had the virtue of scientific and personal humility, so often misunderstood in post-romantic thought as self-abnegation (McCloskey 2006). He worked "not for the making of personal reputation, but because he felt a job wanted doing and was therefore worth doing well" (E. Pearson 1939, 249). In the rough and tumble politics of the academy, and therefore in business and agriculture and law, a humble brewer of Guinness lost out to a very forceful eugenicist.

As Fisher himself said, "The History of Science has suffered greatly from the use by teachers of second-hand material. . . . A first-hand study is always instructive, and often . . . full of surprises" (quoted in Mendel 1955, 6). Unless you understand the first-hand history you are going to continue thinking—as you do if you are a sizeless scientist and as we once did ourselves—that there must be *some* argument for the 5 percent philosophy. You will suppose that 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 must have it wrong.

But when you see how Fisher and his immediate followers achieved their sad victory we think you will change your mind. As Friedrich A. Hayek wrote in 1952:

> The paradoxical aspect of it, however, is . . . that those who by the scientistic prejudice are led to approach social phenomena in this manner [e.g., accept the phenomena if statistically significant, otherwise reject] are induced, by their very anxiety to avoid all merely subjective elements and to confine themselves to "objective facts," to commit the mistake they are most anxious to avoid, namely that of treating as facts what are no more than vague popular theories. They thus become, when they least expect it, the victims of . . . [what Whitehead called] the "fallacy of misplaced concreteness." (54)

We think the now sizeless scientists can fulfill the promise of the nineteenth-century quantitative revolution and become rationally anthropometric, biometric, cliometric, econometric, psychometric, sociometric, technometric, and pharmacogenomic. But they will need to get back to questions of *how much* and *who cares?* They will have to shift attention away from their α-selves and toward their β-selves. Statistical scientists share a common intellectual descent with β-Gosset. But most do not realize that α-Fisher's methods are a mutation because in this instance of the history of science it was the wiser teacher who was spurned by a greedy apprentice. Vague popular theories inherited from the flawed Fisher, such as that *t*-statistics are objective evidence of the existence of an effect, or that the R^2 possesses a substance-independent scale on which a model is said to be significant or not, or—to say it more generally—that fit is the same thing as importance, will have to go.