Contents

Preface xv
Acknowledgments xix

A Significant Problem 1

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, environmental quality, and even life. The substitution we are worrying about here is called "statistical significance"—a qualitative, philosophical rule that has substituted for a quantitative, scientific magnitude and judgment.

1. Dieting "Significance" and the Case of Vioxx 23

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 he meant "low sampling error by an arbitrary standard of variance." Forgetting after Fisher what is known in statistics as a "minimax strategy," or other "loss function," 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 33

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 42

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, objective, and the spins of a roulette wheel. These are no more relevant to the task of finding out the sizes and properties of viruses 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: β-Importance vs. α-"Significance" 57

The most popular test was invented, we've noted, by Gosset, better known by his pen name "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 beer at a good 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 subjectivity-phobic scientists 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 listened to Fisher.

5. A Lot Can Go Wrong in the Use of Significance Tests in Economics 62

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 percent because 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 in economics articles 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 74

We did not study the scientific writings of amateurs. On the contrary, we studied the *American Economic Review* (known to its friends as the *AER*), a leading journal of economics. With questionnaire in hand we read every full-length article it published that used a test of statistical significance from January 1980 to December 1989. As we expected, in the 1980s more than 70 percent of the articles made the significant mistake of R. A. Fisher.

7. Is Economic Practice Improving? 79

We published our article 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 published in the *AER* in the next decade, the 1990s. Sadly, our colleagues were again mistaken. Since the 1980s the practice in important respects got worse, not better. About 80 percent of the articles made the mistaken Fisherian substitution, failing to examine the magnitudes of their results. And less than 10 percent 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? 89

Does globalization hurt the poor, does the minimum wage increase unemployment, does world money cause 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 98

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 significance error was 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 106

New assistant professors are not to blame. Look rather at the report card of their teachers and editors and referees—notwithstanding cries of anguish from the wise 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 123

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 percent philosophy was eventually codified by the *Publication Manual of the American Psychological Association,* which dictated the erroneous method worldwide to thousands of academic journals in psychology, education, and related sciences, including forensics.

12. Psychometrics Lacks Power 131

"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 seventy-thousand-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 140

Psychologists and economists have said for decades that people are "Bayesian learners" or "Neyman-Pearson signal detectors." We learn by doing and 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 overdiagnosis of schizophrenia.

14. Medicine Seeks a Magic Pill 154

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 165

Some medical editors have battled against the 5 percent 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,* forced change in his journal. But only his journal. 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 faith 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 176

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 percent. An eleventh study published in the *New England Journal of Medicine* dismissed the drug. The dismissal was based not on effect size bounded by confidence intervals based on 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 187

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 3σ as Definitely Significant": Pearson's Rule 193

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 and one of the founding fathers of neopositivism and 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 203

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 207

Gosset revolutionized statistics in 1908 with two articles 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 and the economic design of experiments. He conceived in 1926 the ideas if not the words of "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 pennywhistle, repaired entire, functioning fishing boats with a penknife, and—though a great scientist—was for thirty-eight years a businessman brewing 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 214

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.

22. How the Wasp Stung the Bee and Took over Some Sciences 227

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 many 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.

23. Eighty Years of Trained Incapacity: How Such a Thing Could Happen 238

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.

24. What to Do 245

What, 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. But it has the merit at least of being relevant to science, business, and life. The Fisherian procedures are not.

A Reader's Guide 253
Notes 255
Works Cited 265
Index 289