SCIENCE AFTER GOSSET, FISHER, AND MATRIXX V. SIRACUSANO

THE CULT OF STATISTICAL SIGNIFICANCE

By Stephen T. Ziliak
Roosevelt University Chicago
Sorbonne IAE de Paris
16th November 2011
Mindless “Significance Testing” is Costing Us Jobs, Justice, & Lives

The test of statistical significance is the most important technique in the empirical branches of the life and human sciences, economics to medicine - and it is broken

The main problem?
80-to-90% of scientists don’t “test for” or “estimate” what we want, which is: Oomph and its odds (but Oomph, especially)
Examples of Oomphless Science in Economics, Government, & Medicine

Two diet pills, Oomph v. Precision which pill for Mom?

Zero black unemployment rates why can’t we find them?

4,953+ cases of Vioxx why insist on 19-to-1 odds?

9-0 Supreme Court Decision, Matrixx v. Siracusano Zicam and loss of smell

Matrixx v. Siracusano (March 22, 2011) presented the Supreme Court of the United States with the question whether a plaintiff can file a claim of securities fraud against a company which fails to warn investors about adverse effects that are not statistically significant.

Matrixx Initiatives—the maker of Zicam—claimed not, arguing that a bright-line rule of statistical significance is necessary to establish causation.

The question was considered and decided by the Supreme Court in light of rule §10(b) of the Securities Exchange Act of 1934 as amended by §10b-5.
“Matrixx’s argument rests on the premise that statistical significance is the only reliable indication of causation. This premise is flawed.”

[Supreme Court Decision, Matrixx v. Siracusano, p. 11]

“We conclude that the materiality of adverse event reports cannot be reduced to a bright-line rule.

“Although in many cases reasonable investors would not consider reports of adverse events to be material information, respondents have alleged facts plausibly suggesting that reasonable investors would have viewed these particular reports as material.”
The “Significance” Mistake did not begin with “Student” – the inventor of Student’s $t$ test

A chemist by training, William Sealy Gosset (1876-1937), aka “Student”
Learned statistics on his own—
To solve inference problems in the Main & Experimental divisions of Guinness’s Brewery, Dublin

Copyright: The Galton Laboratory, University College London
Student was a Great Experimentalist
[And a business person, too]

He invented or inspired half of modern statistics: small sample theory, power of the test, balanced and random designs of experiments, Monte Carlo, to name a few.

He co-invented 3 barley varieties grown by the 1920s on 5 million acres, feeding breakfast eaters, beer drinkers, & other wild beasts.

And he served at Guinness’s as:
Apprentice Brewer (1899-1906),
Head Experimental Brewer (1907-1935),
Head Statistician (c. 1922-35), and Head Brewer (Park Royal & Dublin, 1935-37)
Our Father of Guinnessometrics took an Economic and Contextual Approach to the Logic of Uncertainty, 1904 to 1937

“Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment.

What the odds should be depends:

1. On the degree of accuracy the experiment allows

2. On the importance of the issues at stake” (W. S. Gosset, 1904)
Gosset’s report focused on MALT EXTRACT, measured in degrees saccharine per barrel of 168 lbs. malt.

133° saccharine gave the targeted level of alcohol (excise tax was proportionate to alcohol level)

± .5° was an error that beer drinkers and tax payers could swallow

Source: Ziliak, “Guinnessometrics” Jnl of Econ. Perspectives (Fall 2008)
“It might be maintained that malt extract should be [estimated] within ± .5˚ of the true result with a probability of 10 to 1.” He calculated extract means from a series of trials produced in the Main and Experimental Breweries.

Given the small samples he calculated the odds of observing the stipulated accuracy:

“Odds in favour of smaller error than .5

2 observations 4:1
3 “ “ 7:1
4 “ “ 12:1
5 “ “ 19:1
82 “ “ practically infinite”

CONCLUSION n =4 does the trick.
Sort of. “How, in general, should one set the odds with small samples?”
Fisher’s Campaign for a 5% Rule for Student’s Test is Causing More than Headaches

Statistical Methods for Research Workers (1925)

Design of Experiments (1935)

Statistical Methods and Scientific Inference (1955/1956)

Statistical Tables for Bio., Agri., and Medical Res. (with Yates, 1938)

And in scores of articles, letters, and speeches
“The value for which \( P = 0.05 \), or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.”
“Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and *ignore entirely all results which fail to reach this level.*”
“It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results.”
Student told Karl Pearson in 1905:

“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.”

Source: W. S. Gosset to Karl Pearson, c. April 1905, in E. S. Pearson 1939, pp. 215-216.
Student, the Head Brewer of Guinness, did not base decisions on statistical significance

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

Source: “Student” to E. S. Pearson 1937, in Pearson 1939, p. 244. (Pearson was the editor of Biometrika in the era before David Cox)
Fisher Pushed for a Non-Economic, Non-Bayesian Approach and The Cult Listened

Fisher 1955, “Statistical Methods and Scientific Induction,” p. 75:

“Finally, in inductive inference we introduce no cost functions for faulty judgments . . . In fact, scientific research is not geared to maximize the profits of any particular organization . . . We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any currency.”
Matrixx v. Siracusano (March 22, 2011) presented the Supreme Court of the United States with the question whether a plaintiff can file a claim of securities fraud against a company which fails to warn investors about adverse effects that are not statistically significant.

Matrixx Initiatives—the maker of Zicam—claimed not, arguing that a bright-line rule of statistical significance is necessary to establish causation.

The question was considered and decided by the Supreme Court in light of rule §10(b) of the Securities Exchange Act of 1934 as amended by §10b-5.
“Matrixx’s argument rests on the premise that statistical significance is the only reliable indication of causation. This premise is flawed.”

[Supreme Court Decision, Matrixx v. Siracusano, p. 11]

“We conclude that the materiality of adverse event reports cannot be reduced to a bright-line rule.

“Although in many cases reasonable investors would not consider reports of adverse events to be material information, respondents have alleged facts plausibly suggesting that reasonable investors would have viewed these particular reports as material”
“Statistical importance [that is, statistical significance] can’t be a measure because it depends on the nature of the study”.

Justice Kagan, Oral Arguments, Matrixx v. Siracusano (the Zicam case)

“If I heard that, ten people went blind, three used it in one eye, [and] three went blind in that eye, I'd stop using the product.”

“And if I were holding stock in that company, I would sell the stock.”
“This statistical significance always works and always doesn’t work.”

Justice Breyer, Oral Arguments, Matrixx v. Siracusano, Jan. 2011
Chief Justice Roberts appealed to the “total mix of information” required by the “reasonable investor”, as in Basic v. Levinson (1976)

CHIEF JUSTICE ROBERTS:
“An investor is going to worry about the fact that thousands of unreasonable investors are going to dump their Matrixx stock. "

Justice Roberts, Oral Arguments, Matrixx v. Siracusano
JUSTICE SCALIA: “Mr. Shah, what do you think about Satan? “

MR. SHAH: “Let me try to unpack the satanic connection hypotheticals a little bit. “

“Something more is needed,” Justice Sotomayor wrote. But what? And how can statisticians, editors, grantors, and policy makers help?

The Court asks for something more about the “source, content, and context” of information disclosed to investors and the S.E.C.

Justice Sotomayor, author, Matrixx v. Siracusano, Supreme Court of the United States, March, 2011, p. 15
What the Supreme Court did not say is that statistical significance gives us the wrong information—false hope and skepticism

(1) The “Size Matters/How Much” Question cannot be answered by statistical significance—

Ziliak and McCloskey find that evaluation of regression coefficients is eschewed in 8 or 9 of every 10 articles published in leading journals of science—actuaries included

(2) The Probability of a Hypothesis (degree of belief) cannot be revealed by a Fisher test of significance—the fallacy of the transposed conditional.
Student’s economic approach was advanced by Harold Jeffreys’s *Theory of Probability* (1939) & Egon S. Pearson (1938), for example.

Student’s sampling distributions and tests of significance were based on actual repetitions of experiments, not the frequentists’ imaginary
If we stop doing 5% philosophy, what will we do?

Ask to see the Oomph, loss function, and power of a scientific model, variate or stimulus package.

Demand to know the expected value, subject to “real error”, not mere statistical significance.

Listen to a Guinness brewer & don’t settle for small beer – go for substantive significance!
Most textbooks have to be revised. We should teach less Fisher, more Student. Student offers a way forward.
And, of course, write more Haiku

Statistical fit:
epistemological
strangling, of *wit*!

Little *p*-value
What are you trying to say
of significance?

The contents of these slides—including the quotes—are from S.T. Ziliak’s and D.N. McCloskey’s The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (2008, University of Michigan Press); Ziliak’s “Guinnessometrics: The Economic Foundation of ‘Student’s’ t,” Journal of Economic Perspectives (Fall 2008); and Ziliak’s “Matrixx v. Siracusano and Student v. Fisher,” Significance 8 (3), 2011, Royal Statistical Society and ASA.

Copyright 2011
Stephen T. Ziliak