Matrixx *v* Syracusano and Student *v* Fisher

Statistical significance on trial

"Statistical significance is the only reliable evidence of causation? That premise is flawed." So ruled the US Supreme Court this year in a judgment of huge significance. As a drug company learns that it must disclose even statistically insignificant side effects, **Stephen T. Ziliak** congratulates the judges, and two giants of statistics fight on from their graves.

The mistaken equation was evident to Student from the beginning.

"Statistical significance is easily mistaken for evidence of a causal or important effect, when there is none." That was what the man we know as Student told his managing director at Guinness's Brewery sometime around 1904. Similarly, a lack of statistical significance – statistical *in*significance – is easily though often mistakenly said to show a lack of cause or effect when in truth there is one, the future head brewer and inventor of Student's test of significance also said[1,2].

The economic approach to the logic of uncertainty – pioneered by Student at Guinness – can help to minimise what we might as well call the "significance mistake". It is a mistake that is still being made. Bruno de Finetti said in the 1970s that "The economic approach seems (if not rejected owing to aristocratic or puritanic taboos) the only device apt to distinguish neatly what is or is not contradictory in the logic of uncertainty".[3] The test of significance cannot do it.

Enis seditem olore, cum dolestrume di te poriori rehentotat. Nus, conserita sam, quate

Indeed, when the London economist and statistician Francis Y. Edgeworth used the word "significance" early on, in an 1885 article he prepared for the Royal Statistical Society[4], his attention was largely focused on economic interpretations[5]. Edgeworth wanted to understand the economic reasons for the magnitude, the effect size. He wanted to know the practical importance of coefficients he estimated from bee and wasp migration data collected while hanging around Hampstead Heath, his local park. For what percentage of their time were they away from their nests foraging? Edgeworth thought to control for variance (of this quantity?), and showed others how to do the same. But his methods did not give undue weight to the denominator, that is, to the amount of probable dispersion around magnitudes of economic interest. He did not make those probabilities his criteria of judgment. Others have not been as wise as Edgeworth, or as Student.

"Student" was actually William S. Gosset. His employers, Guinness, would not let him publish under his own name so he used that *nom de plume* instead. We owe him a great deal, especially for hand-calculating and publishing the first four versions of Student's *Small Sample Table of Statistical Significance*, between 1908 and 1925. Those tables are still used today (in digital form, of course) to establish the statistical significance of experimental results. Yet despite that great work, he did not himself entrust judgments to statistical significance – not at any level[5].

He explained in a letter of 1905 to Karl Pearson that fixing a conventional level of significance – 95% certainty, or 99% certainty – of a relationship is not a reasonable choice for industry. He observed that the degree of probability is merely one element of the expected

profit and loss function or of what he called "pecuniary advantage" and "cost":

> When I first reported on the subject[6], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment.[7,8]

Profitable decisions, Student told Pearson, are obstructed, not enabled, by an arbitrary rule about the level of significance.

Strangely, this and other valuable knowledge about significance has not much prevailed in science or law during the past century. Ziliak and McCloskey show that 8 or 9 out of every 10 articles published in the leading journals of science commit the significance mistake – equating significance with a real and important practical effect while at the same time equating insignificance with chance, randomness, and no association or causal effect at all[5].

By the 1930s the null hypothesis significance test procedure at the 5% level was considered the height of advance in fields such as economics, psychology, and medicine, "that these studies might be raised to the ranks of sciences", in the words of Ronald A. Fisher[9].

"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", Fisher said again in 1926, and then again and again, in different versions, for most of his nearly fifty year career[10].

Student was not impressed. Fisher and he battled about significance till the day Student died, in October 1937, just as Student was completing his *Biometrika* article on "Comparison between balanced and random arrangements of field plots", replying to a 1936 attack by Barbacki and Fisher. Yet as Ziliak, McCloskey, and others have shown – and as many people have experienced – the Fisher School got its way, spreading and enforcing artificial rules about the level of significance, randomisation of design, and other tools of science[11].

Until now, it seems. The founding fathers of modern statistics might be a tad surprised to hear (if they could) that the significance debate rages on – recently, in the Supreme Court of the United States.

Now there is a rule of law upholding the wisdom of Student and other critics of significance. On March 22nd, 2011, in *Matrixx Initiatives, Inc.* v *Siracusano*, in an important case of securities law the Supreme Court of the United States unanimously rejected a bright-line, one-size-fits-all, hard-and-fast rule of statistical significance[12].

Matrixx reloaded

The case involves a homeopathic medicine called Zicam, a zinc-based common cold remedy produced by Matrixx Initiatives, Inc. When applied through the nose the drug causes some users to experience burning sensations and others to suffer anosmia – that is, the permanent loss of smell[12].

Unt fugia si consero rerumqui blantis a voluptatet laborib erferor iberorepe optae por

Matrixx ignored a number of adverse reports it had received from doctors and users since 1999. One doctor told the company that zinc toxicity was discovered by biologists back in the 1930s. He suggested the company might wish to look into the amount and type of zinc it is putting up people's noses. No reply; no adverse effect disclosures. When a doctor appeared on *Good Morning America* in 2004 and told the untold story of Zicam, Matrixx stock price plummeted. Again the company hid, this time behind the argument that the adverse effects were – wait for it – "not statistically significant". The company assured investors that revenue was expected to grow vastly – by "50 and then 80 percent"[12] – for the current hundred million dollar a year seller, despite the company's knowledge of zinc toxicity and the adverse effect reports. Following a warning from the Food and Drug Administration, the nasal spray and swab forms of Zicam were taken off the market. Over 200 Zicam users have filed lawsuits against Matrixx, with more lawsuits pending. Mr Siracusano, however, was not a user no longer able to enjoy the gorgeous scent of roses; he was suing on behalf of investors in Matrixx who believe that they were misled.

The company did not actually conduct research on the active zinc ingredient, a fact which seemed to irritate the Justices. Instead the Zicam maker invoked the old fallacy which Student and other statisticians warned about, asserting incorrectly that the loss of smell was not "important" because it was not "statistically significant". Again the court was not amused. In the oral argument on January 10th, 2011, Justice Sotomayor chastised petitioners' counsel, Mr Hacker, for neglecting technical briefs on the subject that had been authored and filed by *amici* (friends) of the court, many of whom, the Justice said, "did a wonderful job". (Modesty should forbid, but full disclosure requires, that I state at this point that I was one of the *amici*.)

Investors in Matrixx (respondents) had previously filed suit against the company in a federal district court. Investors told the district court that Matrixx failed to disclose relevant information about adverse effects that the company had received from expert nose doctors. The district court dismissed the case on the basis that investors did not prove "materiality", by which that court meant "statistical significance".

The US Court of Appeals for the Ninth Circuit reversed the decision of the district court. The Appeals Court "reasoned that whether facts are statistically significant, and thus, material, is a question of fact that should ordinarily be left to the trier of fact – usually the jury".

But the Supreme Court of the United States disagreed with the significance-based definition of materiality invoked by the district court. The Supreme Court Justices said that the lower court "erred when it took liberties in making that determination on its own".

Matrixx v *Siracusano* presented the Supreme Court with the question whether a plaintiff can file a claim of securities fraud against a company which failed to warn investors about adverse effects that are not statistically significant. 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.

The court's unanimous rejection of significance will affect drug supply and demand, securities regulation, liability and, notably, the content and frequency of adverse effect reports disclosed (or not) by drug companies. For example, a pharmaceutical company reporting to the Securities and Exchange Commission

can no longer hide adverse effect reports from investors on the basis that reports are not statistically significant. "The court ruled that the materiality of a pharmaceutical company's non-disclosure of adverse event reports in a securities fraud action does not depend upon whether there is a statistically significant health risk."

"Something more is needed", Justice Sotomayor wrote of adverse effect reporting, and that something more should address the "source, content, and context" of the adverse information[12].

Significance, causality and the reasonable investor

The court invoked the expectations of a reasonable investor. Would an undisclosed adverse effect report be likely to negatively affect the "total mix" of information considered by a reasonable investor? If yes, then the report must be disclosed, regardless of statistical significance[12].

The court asked, given that researchers, the FDA, and medical experts do not require statistical significance, why a reasonable investor would insist on statistical significance. Justice Sonia Sotomayor, the author of the opinion, said:

> medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence. … The FDA similarly does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data. In assessing the safety risk posed by a product, the FDA considers factors such as "strength of the association," "temporal relationship of product use and the event," "consistency of findings across available data sources," "evidence of a dose-response for the effect," "biologic plausibility," "seriousness of the event relative to the disease being treated,""potential to mitigate the risk in the population,""feasibility of further study using observational or controlled clinical study designs," and "degree of benefit the product provides, including availability of other therapies." … [The FDA] "does not apply any single metric for determining when additional inquiry or action is necessary".[12]

Likewise during the January 10th, 2011 oral arguments preceding the March decision, the Justices of the Court did not express much regard for statistical significance[13].

To the Matrixx argument that statistical significance set the standard for disclosure, over and above "background noise", Justice Breyer replied to Matrixx representative Mr Hacker (that is his real name, not a plot by me or Charles Dickens):

> JUSTICE BREYER: Oh, no, it can't be. I mean, all right – I'm sorry. I don't mean to take a position yet. But –
>
> (Laughter.)
>
> JUSTICE BREYER: But, look – I mean, Albert Einstein had the theory of relativity without any empirical evidence, okay? So we could get the greatest doctor in the world, and he has dozens of theories, and the theories are very sound, and all that fits in here is an allegation he now has learned that it's the free zinc ion that counts.
>
> MR. HACKER (for Matrixx): But –
>
> JUSTICE BREYER: And that could be devastating to a drug even though there isn't one person yet who has been hurt[13].

To Hacker's argument that statistical significance is the way to reason, Justice Breyer snorted back: "This statistical significance always works or always doesn't work".[13] Justice Sotomayor said in oral argument, citing *amici*, that what counts as "statistical importance can't be a measure because it depends on the nature of the study".[13] Justice Kagan and Justice Ginsberg argued that small numbers of meaningfully large effects can be materially relevant, independent of the level of statistical significance. Justice Kagan considered a couple of situations in which a small number of instances of blindness were known to be associated with the use of a specific contact lens solution. She said that the FDA would not wait around for statistical significance to make a determination or to investigate further into the facts of these tragic black swans.

Chief Justice Roberts sympathised with the informational expectations of a "reasonable investor" and concluded that statistical significance was not necessary for establishing causation or belief in association[13].

"Satan", "psychics", and "barking lunatics" made several cameo appearances during oral arguments about significance. These characters were mentioned and discussed by several of the Justices but no one in the court seemed to know the exact significance or importance of those appearances. Perhaps they had come to a conclusion.

"Matrixx's argument rests on the premise that statistical significance is the only reliable indication of causation. This premise is flawed".[12] "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".[12]

Something more

The Matrixx decision is consistent with the court's earlier rejection of a bright-line rule in a fact-finding and economically important situation. Citing *Basic* v *Levinson* (1976), a case involving a bright-line definition for what is meant by "merger negotiations", Justice Sotomayor reminded the court: "we observed [in *Basic*] that '[a]ny approach that designates a single fact or occurrence as always determinative of an inherently fact-specific finding such as materiality, must necessarily be overinclusive or underinclusive.'"

Consider a scenario where a pill is thought to be quite effective at providing pain relief but at the cost of an increased risk of heart attack for the pill user. Suppose a well-designed experiment is conducted on a sample of adult humans, half taking the drug, the other half taking another, competing drug.

The significance tester – in search of a single, determinative fact – then poses a question: assuming there is no difference between the two pills, assuming they are the same, what is the chance that the data we have on hand, showing some amount of difference, will actually be observed?

If the chance of seeing the adverse effect is less than or equal to 5%, say ($p \leq 0.05$), then the finding is said to be statistically significantly different from the null hypothesis of "no difference between the two drugs", without

saying how much that difference is, or how one should view it. Whereas if the chance rises above 0.05, the finding is said to be insignificant; there is not enough of an indication to report, significance testers and Mr. Hacker of Matrixx have claimed.

But this is false. For example, in the early 2000s Merck got into billions of dollars of trouble with their Rofecoxib (Vioxx) pain pill. Vioxx-takers began to up and die from heart trouble, and not completely for random or unrelated reasons. In a clinical trial the Merck scientists reported that Vioxx takers risk a big adverse effect, death. Yet the *p*-value came in at 0.20 - "statistically insignificant", and by some distance – so the company neglected the adverse outcomes. (There were other problems with the Vioxx study, which we shall not discuss here.)

What the Supreme Court did not go on to say, but could have, is that the test of significance gives us the wrong information, a kind of false hope or undeserved scepticism: it sees things the wrong way up. The significance test calculates – however imperfectly – the probability of seeing the data, assuming that the treatment and control drugs are the same. But that is what Dennis Lindley and I call the "fallacy of the transposed conditional". We have already seen the data; we already know the probability of seeing it: it is 1. The data is the one thing that we already do know, and for certain. What we actually want to know is something quite different: the probability of a hypothesis being true (or at least practically useful), given the data we have. We want to know the probability that the two drugs are different, and by how much, given the available evidence. The significance test – based as it in on the fallacy of the transposed conditional, the trap that Fisher fell into – does not and cannot tell us that probability. The power function, the expected loss function, and many other decision-theoretic and Bayesian methods descending from Student and Jeffreys, now widely available and free on-line, do.

Second – as *amici* Deirdre McCloskey and Stephen Ziliak explain in the brief they filed with the Supreme Court – a "significant" result does not in any way answer the "how much" question, the question of how much or how valuable the difference in magnitude (such as loss of smell or sight) is[14]. If the side effect had been not loss of smell but a slight tingling of the nose which disappeared after 40 seconds and never returned, nobody would have sued anyone. The significant result cannot demonstrate economic, medical, or other importance. In other words, the test of significance does not tell us what we would like to know, which is the probability of detecting a large and practically important difference when the difference is truthfully there; for that, one needs exploratory methods, a power function, an expected loss function, and ideally speaking, a series of independently repeated experiments controlling for random and real error – just as the Guinness brewer and his school have long advised.

If you ask me, Lady Justice deserves a pint.

© iStockphoto.com/ene

References

1. Ziliak, S. T. (2008) Guinnessometrics: The Economic Foundation of "Student's" t. *Journal of Economic Perspectives* **22**(4), 199–216

2. Ziliak, S. T. (2010) The *Validus Medicus* and a new gold standard. *Lancet*, **376**, 324–325.

3. De Finetti, B. (1976) Comments on Savage's "On rereading R. A. Fisher". *Annals of Statistics* **4**(3), 486–487.

4. Edgeworth , F. Y. (1885) Methods of statistics. In *Jubilee Volume of the Statistical Society*, pp. 181–217. London: Royal Statistical Society of Britain.

5. Ziliak, S. T. and McCloskey, D. N. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

6. Student (1904) The application of the "law of error" to the work of the brewery.

7. Letter of William S. Gosset to Karl Pearson, *c*. April 1905, quoted in Egon S. Pearson, "Student" as statistician. *Biometrika*, **30**(3/4), 215–216.

8. Pearson, E. S. and Wishart, J. (eds) (1942) *Student's Collected Papers*. London: Biometrika Office, University College.

9. Fisher, R. A. (1941) *Statistical Methods for Research Workers*, 8th edn. Edinburgh: Oliver and Boyd.

10. Fisher, R. A. (1926). Arrangement of field experiments.*Journal of Ministry of Agriculture*, **33**, 503–513.

11. Ziliak, S.T. (2011) Field experiments in economics: Comment on an article by Levitt and List. Roosevelt University, Department of Economics, May.

12. Supreme Court of the United States, no. 09-1156, Matrixx Inititiaves, Inc., et al., Petitioner v. James Siracusano et al., On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit, March 22, 2011. 25 pp., syllabus.

13. Supreme Court of the United States, no. 09-1156, Matrixx Inititiaves, Inc., et al., Petitioner v. James Siracusano et al., On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit, oral arguments, January 10, 2011. `http://www.supremecourt.gov/oral_arguments/argument_transcripts/09-1156.pdf`

14. McCloskey, D. N. and Ziliak, S. T. with Labaton, E. *et al.* Counsel of Record (ed.), *Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents* (vol. no. 09-1156, pp. 22). Washington DC: Supreme Court of the United States.

Stephen T. Ziliak is a Trustee and Professor of Economics at Roosevelt University, Chicago, and has published extensively on statistical significance and practical importance. He is the author, with Diedre McCloskey, of *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* (University of Michigan Press, 2008).