Review of Social Economy

ISSN: 0034-6764 (Print) 1470-1162 (Online) Journal homepage: http://www.tandfonline.com/loi/rrse20

Statistical significance and scientific misconduct: improving the style of the published research paper

Stephen T. Ziliak

To cite this article: Stephen T. Ziliak (2016) Statistical significance and scientific misconduct: improving the style of the published research paper, Review of Social Economy, 74:1, 83-97

To link to this article: http://dx.doi.org/10.1080/00346764.2016.1150730

Published online: 07 Apr 2016.

REVIEW OF SOCIAL ECONOMY, 2016
VOL. 74, NO. 1, 83–97
http://dx.doi.org/10.1080/00346764.2016.1150730

Statistical significance and scientific misconduct: improving the style of the published research paper

Stephen T. Ziliak

Department of Economics, Roosevelt University, Chicago, IL, USA

ABSTRACT

A science, business, or law that is basing its validity on the level of *p*-values, *t* statistics and other tests of statistical significance is looking less and less relevant and more and more unethical. Today's economist uses a lot of wit putting a clever index of opportunity cost into his models; but then, like the amnesiac, he fails to see opportunity cost in statistical estimates he makes of those same models. Medicine, psychology, pharmacology and other fields are similarly damaged by this fundamental error of science, keeping bad treatments on the market and good ones out. A few small changes to the style of the published research paper using statistical methods can bring large beneficial effects to more than academic research papers. It is suggested that misuse of statistical significance be added to the definition of scientific misconduct currently enforced by the NIH, NSF, Office of Research Integrity and others.

ARTICLE HISTORY Received 22 January 2016; Accepted 22 January 2016

KEYWORDS Research ethics; econometrics; Bayes factors; human values; statistical reporting

Statistics are not lies

The scientific community is charged with, among other things, the task of establishing criteria for summarizing evidence in favor of new hypotheses, for discerning quantitative differences between phenomena of interest – from stimulus packages and star clusters to safe food and drugs – and for guidance in general in making rational decisions or changes of belief under conditions of uncertain variance.

After Laplace and Gauss – and especially after Galton and Pearson – in sciences from physics and economics to biology and medicine, statistical methods – the theory, design, measurement, analysis, and interpretation of models and data – have offered up the best tools of the trade.

CONTACT Stephen T. Ziliak sziliak@roosevelt.edu

That might sound cloying, or even surprising, to non-statistical folk more than a century after Karl Pearson's *The Grammar of Science* (1892). (On the basis of a gift from Francis Galton, Pearson founded at University College London, in 1911, the first Department of Applied Statistics in the English speaking world.) But insiders to the field know what Pearson and Galton knew – that Twain and Disraeli were wrong. It is simply not true that there are "lies, damned lies, and statistics." Or, to be exact, not in that order of ascending blunder. On the contrary, scientists – and most politicians after Stalin – know that it is easier to lie or deceive when *not* using statistics. Statistical methods invite and challenge scientists and citizens alike to consider the quantitative evidence – to estimate the size, variance, and probable relationships between observations.

Common sense confirms that model-based evidence is never the whole of human judgment, and adherents to both Frequentist and Bayesian Schools of statistics agree (for instance, Cox 1986; Gelman 2013; Press 2003; Rosenthal and Rubin 1985; Senn 2001). Some human values affecting judgment – such as mercy or justice in medicine, or fear of loss in cognitive psychology – do not fit snugly into the math. Plus, human knowledge – from big data to small – is limited. Only 10 years ago the profit that could be earned from social media was predicted by few. Innovation is not easily predictable. If it were, it wouldn't be an innovation. But for many phenomena, statistical evidence – when available – contributes a lot.

Statistical significance on trial

Currently there is a big question brewing in science and society regarding the proper role – if any – for one part of statistical analysis: "statistical significance." Recently, a journal of social psychology has discouraged use not only of statistical significance but of inferential statistics in general (Trafimow and Marks 2015). More recently the editors of eight of the leading journals of health economics have joined forces to issue an editorial statement urging increased attention to statistically "insignificant" findings that are nevertheless humanly and economically important to real people and health economics (Editors 2015). Journalists and critics have weighed in, too, raising concerns about the rise of "*p*-hacking" – that is, unscientific searches by scientists hoping to attain high levels of statistical significance merely for the sake of publishing (Bowyer 2015; Nuzzo 2014; Simonsohn *et al.* in press). And in medicine, business, and law, the statistically significantly mishandled Vioxx pill has cost its producer several billions of US dollars in liability payments to Vioxx victims and their families – one of the biggest liability payouts by a single corporation in American economic history (Ziliak and McCloskey 2008: 28–31).

Perhaps the most important change to the thinking about statistical significance is a unanimous 2011 decision by the Supreme Court of the USA (Bialik 2011; Supreme Court of the United States 2011). In a major case of securities

law regarding biomedical companies' public disclosure of adverse events the High Court overturned previous law, deciding 9–0 *against* using statistical significance as a standard for whether or not a firm has to disclose bad news. The Court ruled that statistical significance does not and cannot determine the fact of materiality and importance to lives and fortunes.

What is statistical significance?

Ironically, it's the expert econometrician and professional statistician – not the workaday significance tester – who mostly agrees with the critics. Statistical significance is by itself neither necessary nor sufficient for proving a scientific, commercial, medical, or legal claim. Rational assessment of the probability or likelihood of a hypothesis cannot be derived from statistical methods alone – Bayesian methods included. After all, the choice of loss functions and samples and alternative hypotheses to study requires argumentation and appeals to warrants of belief – both ethical and scientific – that are not contained in the data or estimates. In particular, significance tests using *p* values, *t* statistics, *F* statistics, and confidence intervals do not by themselves provide a sound basis for concluding that an effect is important, that is, scientifically, legally, practically, or economically important. Significance and importance are weakly correlated. Reasoning of values and purposes including, but by no means not limited to, considerations of efficiency and fit beyond those contained in the analyzed data and in conventional statistical models must be used to reach such a conclusion.

Critics are right: too many sciences – and in the USA, the law itself, up until the recent Supreme Court decision on Matrixx v. Siracusano – have fallen into the habit of claiming that statistical significance – or the lack of it – statistical "insignificance" – is a decisive, yes/no deal breaker for purported research findings; it's not. But the mistaken usage comes in numerous forms, from logical to practical, and is similarly distributed across the different sciences (Goodman 2002; Greenland and Senn 2015; Lew 2012s). Ziliak and McCloskey (2008) report from a large survey of science in the twentieth century that 8 or 9 of every 10 articles published in the leading journals of economics, health, medicine, and psychology misused (at a cost) the test of significance. Ioannides (2005) took Freiman and others' (1978) study of "71 'Negative' Trials" to the next level, arguing "why most published research findings are false." Yet, one should not conclude from past blunders, as some scientists have, that inferential statistics should be taken off the table. Or that tests of significance have no place in the ethical-scientific toolbox.

Significance testing does have a place. But admittedly, the old confusion begins with the very word, "significant." It's not like it sounds. "Significant" in the statistical sense does not mean "important." And "insignificant" does not mean "unimportant." This elementary point should be more common knowledge than in reality it is.

Of course some of the observed variance of numerical observations will be random, some systematic. For example, in agricultural economics the estimated level of a farmer's yield per acre when growing barley A will not always surpass the yield of barley B, grown on the same or different fields. Some of the difference in yield will occur for random reasons – from uneven occurrence of rabbit holes or bird attacks, say. The other part of the variance might be explained by systematic reasons – from differences in the yield of A and B caused by genetic or soil differences, or by differential access to water and sun.

Misconduct rising: when Student's *t* became Fisher's *p*

Since Student's (1908) article on "the probable error of a mean" it has proven helpful to estimate how likely the observed differences might be, and by how much. The most commonly used test is called Student's test of statistical significance, or "Student's *t*." The *t*-test produces a *t*-statistic which one may use as an input for judging the likelihood of events. Commonly the test assumes one has a random sample of independent and repeated observations exhibiting a normal "bell shaped" distribution.

The *t*-statistic is a ratio which divides observed mean differences between phenomena under comparison (such as the average difference between barley yields A–B) by a measure of its own variation, called the standard error. By convention, in most of the life and human sciences, if Student's *t* rises to twice the level of observed variation in the data (twice the standard error) then the finding is called "statistically significant;" if not, not.

The "*p* value" estimates something more abstract and difficult to grasp: *p* measures the likelihood that the value of Student's *t*-statistic will be *greater* than the value of t actually observed assuming to be true the assumptions of the model and the stipulated null hypothesis of "no average ~~or median~~ difference" between the things being compared.

Assuming that there is no average difference between the two barley yields – or between two pain relief pills or two star clusters or two investment rates of return – the *p* value calculates the likelihood that we would observe a deviation in the data at least as large as the deviation, measured by Student's *t*-statistic, that is actually revealed. (In many testing situations scientists are comparing more than two things at once.) Prior knowledge does not figure into the calculation of *p* – a fact which both Frequentist and Bayesian Schools lament – one more reason to consider all of the available information, both inside and outside the model. If the calculated value of *p* is low – below the overused 0.05 level, for example – convention claims the data on hand are not consistent with the null hypothesis of "no difference" between the objects being compared and concludes that the result is "statistically significant," that is, statistically significantly different from the null hypothesis; if not, not.

That is, if the *p*-value rises to $p = 0.06$ or 0.20 a convention enforcing the bright line rule of $p < 0.05$ would declare "insignificance" and ignore the finding – however important the effect size might be in other regards. For example, in the clinical trial on Vioxx, the heart attack variable came in at about $p = 0.20$. Following convention, the dangerous side effects were thus neglected by its marketer – an unfortunate fact which was finally brought to the attention of the US Supreme Court (McCloskey and Ziliak 2010: 14–18). The data are too noisy and there might not be an effect different from the null, convention claims. But that is not so.

So why should we measure differences between objects using *p*? By all indications, it appears we should not. As the eminent Bayesian and geo- and astrophysicist Jeffreys ([1939] 1961) observed long ago:

> If P is small, that means that there have been unexpectedly large departures from prediction [under the null hypothesis]. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value [of the test statistic] is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law [or null hypothesis], not against it. The same applies to all the current significance tests based on P integrals.

The trouble in workaday practice spread further when researchers, lawyers, and bureaucrats began to misinterpret larger p values (most commonly $p > 0.05$) as the probability that the null hypothesis is true. But a large-scale replication confirmed by James Berger (2003: 4) and others found that when *p* "is near 0.05, at least 72% and typically over 90%" of tested null hypotheses are "true." Statisticians have long recognized the large and important difference between a *p* value and an actual calculation of the probability of the null hypothesis (Student 1938).

The origin of the controversial *p*-value is itself remarkable, and little known. In 1925, Fisher transposed and inverted "Student's" *t* table to produce what are now called *p*-values (Student 1925: 106; c.f. Fisher [1925] 1928, Appendix). In fact, Fisher's p-values are still found in interval form by consulting "the table of *t*." Exactly why Fisher took liberty to change the table – and thus the method of estimation and interpretation – is a matter of historical debate (Pearson 1990: 83; Ziliak and McCloskey 2008: 227–233).

Keep inferential statistics in the journals

But a recent decision by the editors of *Basic and Applied Social Psychology* (BASP) to "ban all vestiges of the null hypothesis significance test procedure (*p*-values, *t*-values, *F*-values, statements about "significant" differences or lack thereof, and so on)" will do little to resolve these problems, and may create

more serious problems than already exist (Trafimow and Marks 2015). This is like recommending decapitation as a headache remedy: it solves the problem, but kills the patient. The new editorial policy, which has gained international attention, asserts that:

> BASP will require strong descriptive statistics, including effect sizes … Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem.

The community of statisticians lauds any effort to emphasize probable "effect sizes" instead of statistical significance. Science and society want to *how much difference* – how big is big and how small is small? – and what that might mean. From drug trials to dental implants, we want to know the likely magnitude and meaning of observed differences between alternatives, leading to more rational decisions. But "strong descriptive statistics" has no scientific meaning, and "larger sample sizes" invite problems of heterogeneity and correlation that cannot be addressed without resorting to the large body of inferential statistics and tools that the psychology journal is currently trying to ban.

Significance controversy in the past

This is not the first time in history that statistical significance has been on trial. "Significance" was only a partial argument from odds from the beginning, Edgeworth (1885: 208), who coined the term, clearly perceived. Galton and Pearson saw in the test more security than they might have. But by 1905 Student himself – that is William Sealy Gosset aka "Student," the inventor of Student's *t*, and eventual Head Brewer of Guinness – warned in a letter to Karl Pearson about economic and other substantive losses that can be caused by following a bright line rule of statistical significance:

> When I first reported on the subject [of 'The Application of the 'Law of Error' to the Work of the Brewery' (1904)], 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 you, 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*. (quoted in Ziliak 2008: 207)

Student's rejection of a bright-line accept–reject standard was echoed a few years on by Harvard psychologist Boring (1919), warning about the difference between substantive and merely statistical significance in psychological research. Yet, mindless tests and uses of statistical significance raged on, heedless of warnings from its eminent discoverers.

In the 1970s, international committees were formed in medicine, epidemiology, and psychology, to help tame the uncertain beast but their un-concerted

efforts could only effect so much change in the way significance tests were used, and, in the event, not much changed (an exception is the journal *Epidemiology*, edited by Kenneth J. Rothman, which seems to have made empirical progress: Rothman (1998)).

And this is not the first time in history that a scientific journal has tried on its own to ban p-values and other measures of statistical significance. *The New England Journal of Medicine* (in the 1970s), *Epidemiology* and *The American Journal of Public Health* (in the 1990s) and the *Publication Manual of the American Psychological Association* (in fits and starts) have experimented with bans of sorts but the record is – with the possible exception of epidemiology, which became *more* quantitative – not salutary.

In economics, my research on the significance mistake has benefited from published comments by Zellner, Lindley, Granger, Schelling, Elliott, Hymans, Wooldridge, Leamer, Horowitz, Gingerich, Mayer, Hoover, Siegler, Smith, Bergman, Spanos, Harford, and others. Probability theorist Olle Haggstrom, medical researcher Jessica Ancker, philosopher of science Steve Fuller, philosopher Deborah Mayo, historian of science Theodore Porter, and many other scientists and scholars have weighed in. While there is some debate among these authors about the frequency of the significance mistake in economics and other sciences (see Ziliak and McCloskey (2013) for citations) all agree on the fundamental fact of meteorological science: that the deep, dark night of mindless significance testing has to go. Here is a fact rather tough to swallow: today's economist – the trusted guardian of the bottom line – devotes much time and energy and wit to putting a clever index of opportunity cost into his model or game; but then, like the amnesiac or weaker college freshman, he fails to see opportunity cost in statistical estimates and interpretations he makes of those very same models. Medicine, psychology, and pharmacology are, I have shown, similarly harmed by this fundamental error of science, keeping bad treatments and drugs on the market and good ones out.

> Econometrics:
>
> where everything is counted
>
> save the net profit.

Change is coming, and social economists can help. Some already are. In fact in an unprecedented policy-change movement by economics journal editors toward the collective good, the editors of eight of the leading journals of health economics published an "Editorial Statement on Negative Findings" (Editors, Health Economics 2015):

> The Editors of the health economics journals named below believe that … innovative conceptual and methodological approaches … have potential scientific and publication merit regardless of whether such studies' empirical findings do or do not reject null hypotheses that may be specified … Moreover, we believe that this should reduce the incentives to engage in

two forms of behavior that we feel ought to be discouraged in the spirit of scientific advancement:

1. Authors withholding from submission such studies that are otherwise meritorious but whose main empirical findings are highly likely 'negative' (e.g., null hypotheses not rejected).

2. Authors engaging in 'data mining,' 'specification searching,' and other such empirical strategies with the goal of producing results that are ostensibly 'positive' (e.g., null hypotheses reported as rejected).[1]

Statistics within reason

In other words, the type of reasoning and conclusions aided by statistical analysis should occur in a larger context of research purpose, design, and values. Tests should be conducted with added attention to the experimental or observational design; the number of repetitions of the experiment or survey; previous findings and data; controls for confounding, selection bias, multiplicity and naturally benefit and cost. Most errors – indeed, most costly errors – are non-random blunders not captured by the usual test statistics. A study boasting about the size or value of its "randomized" tests is probably blowing smoke (see Ziliak and Teather-Posadas (2016) for a deconstruction of randomized field experiments in development economics, medicine, and industrial organization).

Statistical evidence exists only within a statistical model, a fact that is dangerously absent from the BASP editorial. Different models yield different values when applied to a single set of experimental or observational evidence just as different economic or physical or biological schools of thought lead to different interpretation.

What is to be done?

Alternatives to bright-line rules

Perhaps the most commonly used alternative to classical *t* and *p* is the Bayes factor (Carlin and Louis 2008; Press 2003). For discrete data and simple hypotheses, the Bayes factor represents the ratio between the probability assigned to the data under an alternative hypothesis and the null hypothesis (Johnson 2013). One big advantage of Bayesian analysis is that one can compute the probability of a hypothesis, given the evidence, whereas with the null hypothesis test of significance, measured by a *p* value, one can only speak to the probability of seeing data more extreme than have actually obtained, assuming the null hypothesis of "no difference" (or whatever) to be true. As the Bayesian Jeffreys noted ([1939] 1961: 409):

[1]The journals are: American Journal of Health Economics, European Journal of Health Economics, Forum for Health Economics & Policy, Health Economics Policy and Law, Health Economics Review, Health Economics, International Journal of Health Economics and Management, and Journal of Health Economics.

> Whether statisticians like it or not, their results are used to decide between hypotheses, and it is elementary that if *p* entails *q*, *q* does not necessarily entail *p*. We cannot get from 'the data are unlikely given the hypothesis' to 'the hypothesis is unlikely given the data' without some additional rule of thought. Those that reject inverse probability have to replace it by some circumlocution, which leaves it to the student to spot where the change of data has been slipped in [, in] the hope that it will not be noticed.

Jeffreys went on to explain that if one assigns prior odds between the alternative and null hypotheses, multiplication of the Bayes factor by these prior odds yields the posterior odds between the hypotheses. From the posterior odds between hypotheses, scientists can compute the posterior probability that a null hypothesis is true (or in any case useful or persuasive) relative to an explicit alternative. Classical tests of significance, measured by *t* and *p*, cannot.[2]

Unfortunately, the use of Bayes factors in hypothesis testing is considered by the old school to be controversial – "too subjective," they say. "We don't know our priors" or "How do we know what the alternative hypothesis should be?" Common usage is about the only recommendation for the old debate about subjective vs. objective statistics. As Yogi Berra said, "No one goes there anymore; it's too crowded."

In fact, Johnson (2013) observes that in certain hypothesis tests the alternative hypothesis can be specified so that an equivalence between Bayes factors and *p*-values can be established. Technically speaking, Johnson and others have shown, in one parameter exponential family models in which a point null hypothesis has been specified on the model parameter, specifying the size of the test is equivalent to specifying a rejection threshold for the Bayes factor, provided that it is further assumed that the alternative hypothesis is specified so as to maximize the power of the test. The correspondence between Bayes factors and *p*-values in this setting is just one example of the false demarcation line between objective and subjective.

When an alternative hypothesis exists – and that's the usual situation of science: otherwise, why test? – Bayes factors can be easily reported. Bayes factors permit individual scientists and consumers to use prior information or the principle of insufficient reason together with new evidence to compute the posterior probability that a given hypothesis, *H*, is true (or to repeat, useful or persuasive) based on the prior probability that they assign to each hypothesis. After all – fortunately – we do not have to begin every new observation or experiment from *tabula rasa*; we know some stuff, but we want to know more stuff, however imperfectly. Bayes factors add that information into the calculation comparing the likelihood of alternative hypotheses. For example, Bayes factors provide a clear interpretation of the evidence contained in the data in favor of or against the null: a Bayes factor of 10 simply means that the data were 10 times more

[2]Lavine and Schervich (1999) caution that Bayes factors can sometimes lead to incoherence in the technical statistical sense of that term.

likely under the alternative hypothesis than they were under the null hypothesis. That's a real advance over mushy *p*'s.

Thus the beginning of better scientific conduct entails a synthesis of Bayesian and Frequentist ideas. But prior to that inevitable paradigm shift a number of changes to the style of the scientific paper could go a long way toward overcoming current malpractice. Scientific journals and grantors alike can help by providing incentives to put

(1) *Primary emphasis on estimating and contextualizing the substantive significance of estimates.* For example, published tables and text should emphasize the magnitude and meaning of effect sizes, and the Bayes factor between the null and scientifically or economically significant alternative hypotheses. If *p*, *t*, and *F* are reported, exact values should be given and interpreted – together with estimates of effect size – in terms of odds of occurrence and its substantive meaning on a scale of human and scientific values relevant to the purpose. If the odds of an important effect (a fatal heart attack from taking a drug, for example, or earning a small fortune by selling violins) are 1 in 200, some people – perhaps a lot of people – want or need to hear about it. The health economists are right: there should be no bright line rule that banishes such information, and the Supreme Court agrees.

Focusing on substantive significance and alternative hypotheses does not necessarily entail formal specification of a Bayesian "expected loss function." But more attention should be paid to cost or regret measured by amount of job loss from policy X, or death rates in placebo group Y. Likewise, in sciences such as medicine and economics, joy or net benefit ought to be made explicit, measured on some scale of human values not reducible to any level of statistical significance. Money, health, home runs, and less depression, for example, and by how much.

(2) *Due attention, when relevant, to an expected value interpretation of the recommended hypothesis or "decision" relative to feasible alternatives.* Imagine putting your own skin in the game. If you had to lay down money or health or life to gamble on the basis of your result, how much would you wager, and why? For example, Bayesian statistics allow one to estimate posterior probabilities of gain and loss rather than relying on the old accept/reject routine associated with bright-line rules of statistical significance, heedless of substance. As Savage (1954: 116) noted long ago, "The cost of an observation in utility may be negative as well as zero or positive; witness the cook that tastes the broth." Martha Stewart would agree.

(3) *Explicit discussion and results distinguishing random from real error.* For example, in a small sample analysis, a brewer may wish to know with 10 to 1 or better odds how many samples of wort he needs to mix to

be confident that the saccharine level of the beer stays within 0.5° of the 133° standard he is targeting. The example is "Student's": brewing over 100 million gallons of Guinness stout per annum, "Student" and Guinness stakeholders needed to know (Ziliak 2008: 206). "Real" errors in this context include uneven temperature changes, heterogeneous barley malt, and mismeasurement of saccharine levels – adding up to more error than is allegedly described by *p* or *t*.

(4) *Easy to read visualization of model and coefficient uncertainty.* From farming to pharmaceuticals, we want to know what the entire distribution looks like from the point of view of oomph and precision, effect size and uncertainty. Not just the point mean or median, with a superscript of asterisks declaring "significant" or "highly" so. Remember Stephen Jay Gould's far-above-the-median experience with surviving stomach cancer, discussed in his essay "The Median Isn't the Message" (Gould 1985). Gould's doctor cited a median survival time from diagnosis of about 8 months; but the prolific scholar and writer lived and worked for another 22 years! A study by Soyer and Hogarth (2011) tested the predictive ability of more than two hundred econometricians using linear models. Prediction was most accurate when the experts were only given a theoretical regression line and scatter plot of data. Take away the plots and their ability to relate model error to levels of the dependent variable fell dramatically. For novice and seasoned alike, the books by Tufte are illuminating.

(5) *Alternative tabular presentations of coefficient uncertainty.* Power, Type II error, and Bayes factors, for example. Many years ago Freiman *et al.* (1978) published in the *New England Journal of Medicine* a study entitled "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial." The abstract reads:

> Seventy-one 'negative' randomized control trials were re-examined to determine if the investigators had studied large enough samples to give a high probability (>0.90) of detecting a 25 per cent and 50 per cent therapeutic improvement in the response. *Sixty-seven of the trials had a greater than 10 per cent risk of missing a true 25 per cent therapeutic improvement,* and with the same risk, *50 of the trials could have missed a 50 per cent improvement.* Estimates of 90 per cent confidence intervals for the true improvement in each trial showed that in 57 of these 'negative' trials, a potential 25 per cent improvement was possible, and 34 of the trials showed a potential 50 per cent improvement.

And let us all keep in mind:

(6) *Tests of significance on a single set of data are nearly valueless.* Most tests of significance (*t, p, F*, and the others) and their relevant tables of distribution are designed for use on independent and identically repeated experiments or surveys. No repetition means in effect that you are running a test on $n = 1$. Repetitions are also the best way to control for non-random, systematic errors and violations of assumptions of the model that produces the reported results.

(7) *At best, t- and p-values can suggest reasonable cause for further exploration of the apparent result.* But reasonable cause can originate from belief, too. Whether to continue the inquiry or belief depends on other elements of the decision function, such as in medicine efficacy, profitability, and marketability. Neyman and Pearson contrasted the costs of convicting the innocent and liberating the guilty. Test statistics are not able to answer those size matters/how much?-questions demanded by business, government, and other scientists. *T* and *p* offer no currency for judgment measured by lives or money or rain forest saved. Thus test statistics alone cannot be used to assess the value of the observed result, "significant" or not. The test statistic is not a decision-maker but the scientist, profit seeker or other human observer is.

Getting the incentives right: scientific misconduct and statistical significance

Several of the major institutions for the advancement of science in the USA – from the National Institutes of Health and National Science Foundation to the American Association for the Advancement of Science itself – have sought to define and to enforce national standards for research integrity and ethical scientific conduct. Fabrication or falsification of data, deceitful manipulation, and plagiarism are the most commonly cited forms of misconduct named and pursued. Although gross misuse of statistical significance has led to approval of faulty medical therapies which cause harm to real people – and, in some cases, such as the Vioxx debacle, even death – the scientific community has not added misuse of statistical significance to the list of scientific misconduct. That is a peculiar inconsistency given that researchers engaged in similar types of manipulation or questionable research practices have been penalized by those same agencies. For example, a University of Oregon researcher was recently charged and penalized by the US Department of Health and Human Services for publishing "knowingly falsified data by removing outlier values or replacing outliers with mean values to produce results that conform to predictions" (Office of Research Integrity 2015).

Misuse of statistical significance fabricates results in a similar manner and not only by dropping "insignificant" adverse results in the high-pressure drug

industry. The significance mistake is undesirable, inefficient, and, in most cases – philosophers agree – unethical (see Ziliak and McCloskey (2016) and Ziliak and Teather-Posadas (2016) for theoretical discussion of ethics in empirical economics including drugs and field experiments in development economics). But the significance mistake seems to be outside the bounds of the current definition of scientific misconduct used by government agencies, research universities, and – with the extraordinary exception of the Matrixx v. Siracusano case decided by the US Supreme Court – the legal process when such matters get litigated in a court of law. If we are going to stem and finally stop altogether the widespread misuse of statistical significance we must begin to get the incentives right and in more than improved publication style and journal editorial policy.

Acknowledgments

The article began as part of a keynote address on the use and misuse of randomization and statistical significance in economics and medicine, presented by the author at the International Workshop on Scientific Misconduct and Research Ethics in Economics, Izmir, Turkey, 2014. A previous version was drafted as a discussion paper for the American Statistical Association Ad Hoc Committee on P Values and Statistical Significance. The author wishes to thank John Mullahy and Workshop and Committee participants, especially Naomi Altman, Brad Carlin, Erwin Dekker, Wilfred Dolfsma, Val Johnson, Michael Lavine, Michael Lew, Ioana Negru, Ron Wasserstein, James Wible and Altug Yalcintas for helpful comments and suggestions. Any errors are my own.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Berger, J. (2003) "Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?," *Statistical Science* 18(1): 1–32.

Bialik, C. (2011) "Making a Stat Less Significant," *The Wall Street Journal*, April 11, The Numbers Guy. http://www.wsj.com/articles/SB10001424052748703712504576235683249040812

Boring, E. (1919) "Mathematical vs. Scientific significance," *Psychological Bulletin* 16(10): 335–338.

Bowyer, J. (2015) "Why Most of the Studies You Read About Are Wrong," *Forbes*, December 30, Arts and Letters. http://www.forbes.com/sites/jerrybowyer/2015/12/30/why-most-of-the-studies-you-read-about-are-wrong/

Carlin, B. and Louis, T. (2008) *Bayes and Empirical Bayes Methods for Empirical Analysis*, 3rd Rev. ed, London: Chapman and Hall/CRC Press.

Cox, D. (1986) "Statistical Significance Tests," *British Journal of Clinical Pharmacology* 14: 325–331.

Edgeworth, F. Y. (1885) "Methods of Statistics," *Jubilee Volume of the Statistical Society*, 181–217, London: Royal Statistical Society of Britain.

Editors (2015) "Editorial Statement on Negative Findings," *Health Economics* 24(5): 505.

Fisher, R. A. ([1925] 1928) *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.
Freiman, J., Chalmers, T., Smith, H., et al. (1978) "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial," *New England Journal of Medicine* 299: 690–694.
Gelman, A. (2013) "P Values and Statistical Practice," *Epidemiology* 24(1): 69–72.
Goodman, S. (2002) "A Comment on Replication, P Values, and Evidence," *Statistics in Medicine* 11: 875–879.
Gould, S. J. (1985) "The Median Isn't the Message," *Discover* 6: 40–42.
Greenland, S. and Senn S. (2015) "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations," Working paper. Department of Epidemiology and Department of Statistics, University of California, Los Angeles, August 27.
Ioannides, J. (2005) "Why Most Published Research Findings Are False," *PLoS Medicine* 2(8): 696–701.
Jeffreys, H. ([1939] 1961) *Theories of Probability*, Oxford: Oxford University Press.
Johnson, V. (2013) "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences* 110(48): 19313–19317.
Lavine, M. and Schervish, M. (1999) "Bayes Factors: What They Are and What They Are Not," *The American Statistician* 53(2): 119–122.
Lew, M. (2012) "Bad Statistical Practice in Pharmacology (And Other Basic Biomedical Disciplines): You Probably Don't Know P," *British Journal of Pharmacology* 166: 1559–1567.
McCloskey, D. and Ziliak, S. (2010) *Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents, Matrixx Initiatives Inc. et al. v. Siracusano et al*, Washington, DC: Supreme Court of the United States. Edward Labaton et al, Counsel of Record(eds), Vol. 09-1156, pp. 22.
Nuzzo, R. (2014) "Scientific Method: Statistical Errors," *Nature* 506(7487): 150–152, News Feature.
Office of Research Integrity. (2015) *Findings of Research Misconduct*. Washington, DC: U.S Department of Health and Human Services, Office of the Secretary. https://ori.hhs.gov/content/case-summary-anderson-david
Pearson, E. S. (1990) *Student: A Statistical Biography of William Sealy Gosset*, in R.L. Plackett and G.A. Barnard (eds). Oxford: Clarendon Press.
Pearson, K. (1892) *The Grammar of Science*. London: Walter Scott.
Press, S. J. (2003) *Subjective and Objective Bayesian Statistics*, New York: Wiley.
Rosenthal, R. and Rubin, D. (1985) "Statistical Analysis: Summarizing Evidence Versus Establishing Facts," *Psychological Bulletin* 97(3): 527–529.
Rothman, K. (1998) "Writing for Epidemiology," *Epidemiology* 9(3): 333–337.
Savage, L. (1954) *The Foundations of Statistics*, New York: Dover.
Senn, S. (2001) "Two Cheers for P-values?," *Journal of Epidemiology and Biostatistics* 6(2): 193–204.
Simonsohn, U., Nelson, L. and Simmons, J. (in press) "P-curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*.
Soyer, E. and Hogarth, R. (2011) "The Illusion of Predictability: How Regression Statistics Mislead Experts," *International Journal of Forecasting* 28(3): 695–711.
Student. (1925) "New Tables for Testing the Significance of Observations," *Metron* V (3): 105–108.
Student. (1938) "Comparison Between Balanced and Random Arrangements of Field Plots," *Biometrika* 29(3–4): 363–378.
Student [W. S. Gosset] . (1908) "The Probable Error of a Mean," *Biometrika* 6(1): 1–24.

Supreme Court of the United States. (2011) "Matrixx Initiatives, Inc., et al., No. 09-1156, Petitioner v. James Siracusano et al.," *On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit*, March 22nd, 25 pp., syllabus.

Trafimow, D and Marks, M. (2015) "Editorial," *Basic and Applied Social Psychology* 37(1): 1–2.

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

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

Ziliak, S. and McCloskey, D. (2013) "We Agree That Statistical Significance Proves Essentially Nothing: A Rejoinder to Thomas Mayer," *Econ Journal Watch* 10(1): 97–107.

Ziliak, S. and McCloskey, D. (2016) "Lady Justice v. Cult of Statistical Significance: Oomphless Science and the New Rule of Law," in G. DeMartino and D. McCloskey (eds) *Oxford Handbook of Professional Economic Ethics*, Oxford: Oxford University Press, pp. 352–364.

Ziliak, S. and Teather-Posadas, E. (2016) "The Unprincipled Randomization Principle in Economics and Medicine," in G. DeMartino and D. McCloskey (eds) *Oxford Handbook of Professional Economic Ethics*, Oxford: Oxford University Press, pp. 423–452.