Eur J Epidemiol (2009) 24:111–114
DOI 10.1007/s10654-008-9297-4

BOOK REVIEW

Ziliak S T, McCloskey D N. The cult of statistical significance. How the standard error costs us jobs, justice, and lives

Ann arbor. The University of Michigan Press, 2008. 321 pages

Olli S. Miettinen

Published online: 14 November 2008

The authors introduce themselves by saying, "In our scientific work we are quantitative economists and value statistics as a crucial tool" (p. 1).

Their leitmotif in the book I take to be this (p. 2):

> [O]ne part of mathematical statistics has gone terribly wrong, though mostly unnoticed. … [It] seems—unless we and some other observers of mathematical statistics such as … Rothman … are quite mistaken—that reducing scientific problems of testing and measurement and interpretation to one of 'statistical significance,' as some sciences have done for more than eighty years, has been an exceptionally bad idea.

Devotion to and practice of this reductionism evidently is the authors' concept of "the culture of statistical significance." And the intended meaning of "the standard error" in the book's title is implied by such statements as, "Statistical significance … has become the central and standard error of many sciences" (p. 2).

The book's main point, however, is not merely that there is excessive preoccupation with statistical significance at the expense of other, equally or more worthy considerations. A major added, related point is that, "Statistical significance is … misleading at best" (p. xv), an "exceptionally damaging" error in modern science (p. xvi).

These theses are echoed in the words of K. Rothman on the book's back cover:

> … Ziliak and McCloskey show how economists—and other scientists—suffer from a mass delusion about statistical analysis. The quest for statistical significance that pervades science today is a deeply flawed substitute for thoughtful analysis. …

Rothman's point, rephrased, is that the authors have correctly diagnosed a very high endemic prevalence of a serious state of statistical delusion in our community of epidemiological researchers, among others. Himself he presumably takes to be free of this delusion; and in line with this, the book's 15th chapter, among 24, is entitled "Rothman's revolt."

So, something quite shattering is being said about the fundamentals of our epidemiological research, inherently statistical in nature. But, before we can truly awaken to recognition of the delusionary nature that is being imputed to our pervasive quest for statistical significance, we need to be clear on what the authors—and Rothman—regard as thoughtful and thereby sound statistical analysis. What is it said to be about? and what is said to be its essence?

The authors' teaching is this: "Real science depends on size, on magnitude" (p. 5). "Any quantitative science answers how much, or should. … But medicine, economics, and some other sciences have stopped asking how much, especially in their academic, as against their applied, work" (p. 7). "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*" (p. ix).

I disagree with this, and by no means alone, I believe. In real epidemiological science I take our philosophical discipline to require that we regard all associations/relations, whether viewed causally or merely descriptively, to be nil (in the abstract) until we have good reason to believe otherwise. And so, only when we do get to believe otherwise—so commonly on account of statistical

O. S. Miettinen (✉)
McGill University, Montreal, QC, Canada
e-mail: olli.miettinen@mcgill.ca

significance-testing—do we take up quantification as our investigative and/or inferential concern.

To wit, in intervention trials, involving initial equipoise, the accruing evidence is—and must be—used to sequentially test this presumed equivalence rather than to quantify non-equivalence; and only when, if ever, the sequential significance-testing of the empirical deviation from the no-association/no-relation pattern brings the theoretical (abstract) equivalence to question, does the quantitative concern arise.

But, as ethics calls for terminating the trial once equipoise no longer prevails, meaningfully precise quantification of the relative effect(s) generally is not possible in intervention trials.

In general, meaningful quantification—as a matter of suitably narrow confidence interval—presupposes a high degree of statistical significance against the null state, and in studies carefully designed for validity and ethics this rarely is the case. It thus is not that mathematical statistics has misled us to eschew quantification of, notably, potentially very important effects. Rather, we understand the limits of what we can meaningfully do. The null *P*-value and the proximal bound of a two-sided 95% confidence interval are, commonly, quite interchangeable in their inferential burdens, with the distal bound essentially meaningless.

The idea that some how-much question is "central to … Student's *t*" but has subsequently gone ignored is new to me, and so strange that I am quite astounded that a pair of authors presumably quite well-versed in statistics would casually and seemingly seriously present it. But even more astounding is, how rife the book also is with other gross misrepresentations of the ideas in statistics and, specifically, of the way practitioners of significance testing are supposed to think (à la Fisher).

Remarkably, the authors give no express account of their purported Gossetian how-much approach to data-analysis in statistical science. They give mere hints: the involvement of "the loss function way of thinking" (p. 8) and due attention to "power" (p. 69). So the need is to examine an instructive example in the text.

Suitable for this purpose is the discussion concerning the trial in which treatment with Vioxx was compared with its naproxen counterpart. In it, in the context of equal sizes of the two subcohorts, the respective numbers of patients with heart attacks under the treatment were initially reported to have been five and one, but subsequently revised to eight and one.

Judging from this example, central to the authors' favored how-much approach is the concept of "observed effect" (p. 24). For, in this example it allows them to make 'observations' like this: There was "a Vioxx disadvantage of 5 to 1" (p. 29) and, "The damage was actually naproxen takers one victim, Vioxx takers *eight* victims, not five" (pp. 29–30; italics in the original).

'Observed effect' is, however, a contradiction in terms. For effects are not phenomena and they thereby are not observable, not even in principle. They are, instead, noumena [1] and, as such, subject only to be inferred, if even that.

Further on the 'observed effects,' the authors make a point that, I presume, no competent epidemiologic researcher would agree with: As for whether the study was large enough to allow a quantitative judgement about Vioxx use causing heart attacks, they assert that, "What matters is the total sample size, not the rare heart attacks, … " (p. 30). They make no attempt to justify this amazing claim.

"Gosset," they say, "would have rejected the interpretation of the Vioxx scientists and their 'insignificant' 5-to-1 ratio of heart attacks," proceeding to quote Gosset as having written that, "What the odds should be depends: (1) On the degree of accuracy which the nature of the experiment allows, and (2) On the importance of the issues at stake" (p. 30). The meaning of this in the example at issue the authors do not explain (nor do they elsewhere). But they do remark that, "Widows and widowers and sound-thinking scientists are on Gosset's side" (p. 31).

As other examples of the purportedly Gossetian approach—in which odds in favor of the hypothesis were derived as $(1 - P)/P$, where P is the *P*-value as we know it [2]—are equally uninstructive, we need to examine what the authors find so appealing in Rothman's data-analytic teachings and practices.

The backdrop for this is formed by the authors' statements such as, "the value Gosset intended with his test … was its ability to sharpen statements of *substantive* or *economic* significance" and, "If you yourself deal in medicine … Gosset and we would recommend that you focus on *clinical* significance" (pp. 3, 20; italics in the originals).

In perfect accord with this, the authors find that Rothman made, in 1978, "a crushing case for measuring *clinical* significance" (p. 165; italics in the original), and that in his journal *Epidemiology* he wasn't going to accept statements of statistical significance "at all" (p. 167). He was going to prefer "interpretation" based on "careful quantitative consideration of the data," explaining that, "Misleading signals occur when a trivial effect is found to be 'significant,' … or when a strong relation is found 'nonsignificant,' … (p. 167).

In regard to this, I must say, first, that I view as patently illogical the idea that a logical—and indeed preferable—alternative to statistical significance is substantive significance, clinical significance, for example. For, substantive significance has to do with a utility valuation of a

parameter's particular, usually non-null, value, known or conjectured, while statistical significance has to do with something very different: quantification of evidence pertaining to whether any non-null value of the parameter actually obtains (in the abstract).

Next, I take it to be incontrovertible that when the parameter at issue is not yet known to have a value in the non-null range, thoughtful data-analysis includes attention to the possibility that chance alone could account for the empirical association/relation. So, if significance-testing is ruled inadmissible, how is this possibility to be assessed? I, for one, don't know.

Further, it is not that in proper statistical significance-testing an *effect* is found to be, or not to be, significant (statistically, much less substantively). Instead, the *empirical association/relation* may or may not be found to be statistically significant (as evidence against the null).

Finally, I question the premise here, the idea that investigators are to engage in interpretation of the evidence they bring to the fore. Their unique expertise is in the evidence per se, notably in the genesis of the result(s) in study design and the execution of this, while as interpreters of the evidence (in inference) they are quite biased and, thereby, inconsequential [3, 4].

The first example of Rothman's data-analysis has to do with data from a trial, by others, concerning the efficacy of a herbal medication for mild depression. The rate ratio ("relative risk," "risk ratio") for remission was 2.0, with a null *P*-value of 0.14. The authors of that study are said to have "concluded from the *P*-value that [the treatment] is not clinically effective" (p. 183). But then, "Rothman computed a *P*-value function … He shows that another hypothesis, a fantastically beneficial risk ratio, RR = 4.1, shares the same *P*-value, .14, as the null, RR = 0."

The original authors' conclusion was unjustified to be sure. But the rationale for the idea, admittedly less than fully explicit, that "the symmetry of the *P*-function" (p. 184) and Rothman's calculation based on this adds to the evidence against the null escapes me, totally. No attempt was made to justify it.

Another example in the book of Rothman's data-analysis has to do with trials, again by others, of flutamide treatment of prostate-cancer patients. The book's lead-up to this first addresses the initial 10 trials, saying that each of them "showed that [the treatment] can increase the likelihood of patient survival by 12 percent," and that the authors of these 10 trials "had come to an agreement on the effect size" (p. 184). Then, the authors of yet another trial "found a similar effect," but due to lack of statistical significance (in usual terms) they concluded that "there was no 'clinically meaningfull improvement' (pp. 184–185).

Re-analyzing the data from these studies, the book relates, Rothman et alii concluded that the data from the 11th study "only reinforce the findings from the earlier studies that flutamide provides a small clinical benefit"; and they asserted, the book says, that the real lesson from the 11th study is that "one should eschew statistical significance testing and focus on the quantitative measurement of effects" (p. 186).

Looking back at the last two paragraphs from the vantage of what epidemiologic researchers presumably generally know, the following critical observations need to be made: None of the original 10 trials "showed" a 12 percent effect, nor could the authors of these trials have reasonably "come to an agreement on the effect size." By the same token, it is a contravention of familiar principles of clinical trials to say that the 11th trial "found a similar effect," and equally mistaken is the inference of this trial's authors, clearly at variance with Fisher's teachings (pp. 222–223). As for the conclusions by Rothman et alii, there could not have been such "findings" from the earlier studies. And from the 11th trial there is no lesson to be learned; it serves, only, as an example of un-Fisherian interpretation of the *P*-value [4, 5].

Now, rather than eschew statistical significance-testing, let us do it thoughtfully for the first two examples above. As for the Vioxx example, the exact mid-*P* is 0.011 (one-sided), corresponding to a standard-Gaussian $z = 2.3$. This in turn, corresponds to an objective counterpart of the Bayes factor equal to $\exp[(1/2)(2.3^2 - 0.45)] = 11.2$ [3]. Thus, if a reasonable prior probability for Vioxx treatment causing heart attacks was (as high as) 1%, the corresponding posterior probability would reasonably have been $1/(1 + 99/11.2) = 10\%$—no more. From the trial of herbal treatment of mild depression, the null *P*-value of 0.14 corresponded to $z = 1.1$. The prior probability for the treatment's efficacy presumably was about 0.50 (corresponding to equipoise). The corresponding posterior probability thus could have been calculated to be $1/\{1 + 1/\exp[(1/2)(1.1^2 - 0.45)]\} = 59\%$—no more, again.

The lesson from these tests of statistical significance, a bit more thoughtful than the usual, is that one really should eschew quantitative consideration of effects before there is reasonably good reason to believe, per testing of statistical significance, that there *is* an effect as an object for quantification. The measure of statistical significance (of an empirical association/relation) should be, I urge, not the null *P*-value but its corresponding objective counterpart of the Bayes factor [4].

I need to add that when quantification of an effect justifiably is the concern in epidemiological data-analysis, the parameter to be estimated—in the sense of providing an interval estimate—should be something more meaningful than the proportional change in the "likelihood of patient survival, or odds ratio not otherwise specified, as for the efficacy of a treatment for a cancer. Moreover, the

determinant contrast, effect modification, and study domain require much more careful consideration—in addition to the larger amount of information in the study.

All in all, thus, whereas Ziliak and McCloskey severely criticize our statistical understandings and practices, and Rothman endorses this, we epidemiologic researchers actually need not accept the claim that we use statistical significance-testing as a deeply flawed substitute for thoughtful analysis.

But we can do our significance testings a bit more meaningfully, by carrying them a bit further—commonly to show how nonsignificant the evidence against the state of no effect/association/relation actually is.

References

1. Kant I. Critique of pure reason, translated by Meiklejohn JMD. Amherst, NY: Prometheus Books; 1990. p. 2, 156 ff.
2. Student. The probable error of a mean. Biometrika. 1908;6:1–25.
3. Miettinen OS. Evidence in medicine: invited commentary. CMAJ. 1998;158:215–21.
4. Miettinen OS. Up from 'false positives' in genetic—and other—epidemiology. Eur J Epidemiol. 2008;23. doi:10.1007/s10654-008-9295-6.
5. Goodman SN. *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol. 1993;137:485–96.