JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

DANIEL J. DENIS

Reply to comments on « The modern hypothesis testing hybrid : R.A. Fisher's fading influence »

Journal de la société française de statistique, tome 145, n° 4 (2004), p. 65-68

<http://www.numdam.org/item?id=JSFS_2004__145_4_65_0>

REPLY TO COMMENTS ON The Modern Hypothesis Testing Hybrid: R.A. Fisher's Fading Influence

Daniel J. DENIS *

I would like to thank Michel Armatte, Bernard Bru, Michael Friendly, Jeff Gill, Ernest Kwan, Bruno Lecoutre, Marie-Paule Lecoutre, Jacques Poitevineau and Stephen Stigler for their comments on my article. You are leaders in the field, and I have learned a great deal from your insightful remarks. In this response, I follow up and discuss three topics that link the commentaries, and attempt to resolve the debate where appropriate.

1. Experimentally Demonstrable Phenomena

I argued that Fisher would have wanted us to publish both positive (*i.e.*, statistically significant) and negative results. In their comments Lecoutre *et al.* (page 2) argue that Fisher's words were not "an incitement to account for non-significant results, even worse to publish them" implying that Fisher meant non-significant results should be literally *ignored*. The interpretation of Fisher here is rather ambiguous. However, coupled with Fisher's recommendation that "a phenomenon is experimentally demonstrable when he [the experimenter] knows how to design an experiment that will rarely fail to give a significant result" (Fisher, 1966, p. 14), the idea of not accounting for non-significant results, whether in terms of publication or some other record, would seem ill-advised, and not commensurate with Fisher's original model. Surely, Fisher would not have held that we should celebrate a potential Type I error amongst a group of studies that failed to reject the null hypothesis. If we do not account for non-significant results, how else are we to conclude, even provisionally, whether or not a phenomenon is "experimentally demonstrable"? Indeed, the only possible way to have an account of this "Fisherian demonstrability" is to have a ratio of rejected to non-rejected null hypotheses. True that Fisher would not have us thoroughly discuss the non-significant results, he would still have us *account for them*. Hence, we must interpret Fisher's use of the word "ignore" to mean that of all experiments performed, we will not "celebrate" the non-significant results, but will discuss those experiments in which we rejected the null. At minimum, Fisher would have counseled that we at least place a "checkmark" under the column of failed experiments. As Gill notes (page 5): "he [Fisher] seeks to reduce or eliminate discussion of

* University of Montana.

non-significant findings even if they are occasionally relevant". With Gill I concur. Fisher wanted us to *limit* our discussion of non-significant results. But this does not mean Fisher would have had us literally *ignore* the failed experiment. *We cannot and should not interpret Fisher to have suggested that the investigator pretend as if the non-significant experiment never took place.* Even implicitly, the idea of an "experimentally demonstrable phenomenon" suggests some sort of record-keeping of significant versus non-significant results. Stigler notes the inherent difficulty in attempting to publish all non-significant results, and argues (p. 2) that "some means of gate-keeping will always be required". I agree with Stigler, but hold that Fisher's notion of "experimentally demonstrable" implies a *particular* method of gate-keeping, one in which, *at minimum* (*i.e.*, even if we do not publish negatives) we note which experiments do and do not make it "through the gate". Ideally however, I hold that the Fisherian model would still have these negatives (or at least their "checkmarks") published in some form, as to be able to identify experimentally demonstrable phenomena in the long run.

2. Levels of Significance

A second key point that requires clarification and further discussion is Fisherian levels of significance. Lecoutre *et al.*, argue that social scientists should not be *blamed* in citing Fisher for their choice of the 0.05 level of significance. Agreed. However, the point is not so much that social scientists *cite* Fisher in their use of the 0.05 level as it is with the rigid, dogmatic, and regal status that they grant the 0.05 level – that is precisely where they depart wholly from Fisher's "convenient use" recommendation. As expressed so well by Kwan and Friendly, "To many, 0.05 is a definitive cut-off and pending on which side one's p-values fall, it could mean *proof, respectability, publication*, or despairingly, the *lack of*" (Kwan and Friendly, 2005, p. 4). Fisher knew there was nothing sacred or special about the 0.05 level, just as he knew there was nothing special about the 0.049 level of significance. He offered a *guideline*, that is all. Social scientists mistakenly adopted his recommendation as a true *rudimentary foundation* of Fisherian significance testing, and it is in this gross misunderstanding where today's model is not the least Fisherian. For instance, Armatte tells us that in the 1960s, the *Journal of Experimental Psychology* demanded manuscripts for which rejections of the null occurred at $p < 0.01$. He correctly notes that such demands simply served to spread serious misunderstanding about the significance test. Social scientists need to realize that you can reject null hypotheses at a level of 0.051 or 0.052, etc., just as you *can* at 0.050 or less. The probability of the data under the null hypothesis should not imply an *automatic* rejection or non-rejection of the null hypothesis. There are simply too many factors that influence the significance test (*e.g.*, sample size, estimated population variance), to regularly adopt *any kind of strict level of significance*. Fisher knew this. Social scientists, historically, have not.

3. The Rejected Null and the Inferred Alternative

A third point that merits follow-up is Fisher's notion of the alternative hypothesis. Lecoutre *et al.* write that "Fisher's conception puts emphasis on the *rejection* of the null hypothesis, whenever one could expect scientific inference to bring argument in *support* of a research hypothesis as Fisher himself recognized" (Lecoutre *et al.*, p. 2). Stigler writes (p. 1), "A degree of Fisher's resistance to the discussion of alternatives was nonetheless tied to his wish to distance himself from Neyman". How then are we to interpret Fisher's position on alternatives? The safest verdict amid this Fisherian confusion is probably to conclude that Fisher correctly recognized the *limitations of significance testing with regard to inferring a substantive alternative hypothesis.* Contrary to his opponent's "Acceptance Procedures", Fisher knew that a simple rejection of the null hypothesis in no way whatever, on *any statistical grounds*, pointed to the correct substantive alternative, and hence, coupled with dislike for his statistical competitors, he was reluctant to discuss the alternative. Fisher in no way wanted to associate his significance tests with Acceptance Procedures, and in defense of this, probably took too extreme of a position. However, it should be noted that in social sciences especially, while inferences of the statistical alternative are relatively straightforward, an inference of the correct *substantive* alternative is usually extremely difficult (Bolles, 1962). What is more, this difficulty is hardly a statistical matter (Denis, 2001). See the excellent comments by Armatte and Bru for a few historical examples of the difficulties inherent with conducting significance tests to advance substantive theory.

Where Fisher stood exactly with regard to the alternative hypothesis is difficult to know for certain. At minimum, we can say that in the Fisherian significance testing paradigm, a rejected null hypothesis does not imply an unequivocal inference of the *substantive* alternative hypothesis. Indeed, as correctly pointed out by Lecoutre *et al.*, a rejected null suggests an inference of its complement. However, it should be noted that this complement to the null hypothesis is merely a statement of "not the null", and is completely segregated from *scientific inference.* Inferring the correct substantive alternative is where the job of statistical inference ends and where the job of scientific inference begins. Today's social scientist too often and mistakenly equates a rejected null with a substantive alternative hypothesis. As argued, this is hardly Fisherian.

4. Final Comments

Discussions of hypothesis testing in the social sciences always bring about debate. Indeed, just recently, as a member of the audience at a conference presentation on the teaching of statistics, I reminded the group that although we were discussing how best to teach null hypothesis significance testing, we should actually be teaching (and doing) Bayesian statistics. Stopping a bit short of Gill's words, "The null hypothesis significance test (NHST) should not even exist, much less thrive as the dominant method for presenting

statistical evidence in the social sciences. It is intellectually bankrupt and deeply flawed on logical and practical grounds" (Gill's comments, p. 1), I nonetheless reminded the audience of the many years of well founded criticism directed at NHST, and how most exceptional methodologists (*e.g.*, Bakan, 1966; Cohen, 1994) have long recommended the Bayesian alternative. Of course, my statement was met with the counter-argument that Bayesian statistics must be further evaluated on their own merit, *prior* (yes, pun intended) to considering them as a replacement to NHST. Hence, the debate continues. Indeed, there are some (*e.g.*, Macdonald, 1997) who have recently argued in support of the Fisherian model, and lay blame with Neyman and Pearson for the confusion surrounding today's model of inference. Regardless of which position you choose to defend, it is hoped that the present review of Fisherian significance testing has served to put some of the contributions of R.A. Fisher in their historical place, so that educators and practitioners of statistics may interpret today's hybridized model with care, and perhaps most importantly, with a critical, if not skeptical eye.

Additional References

BAKAN D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.

BOLLES R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, 11, 639-645.

COHEN J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

DENIS D. (2001). Inferring the alternative hypothesis: Risky business. *Theory & Science, 2*, 1.
http://theoryandscience.icaap.org/content/vol002.001/03denis.html.
Also available on-line (http://htpprints.yorku.ca/).

FISHER R. A. (1966). *The design of experiments.* New York: Hafner Publishing Company.

GILL J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Chapman & Hall/CRC. New York.

MACDONALD R. R. (1997). On statistical testing in psychology. *British Journal of Psychology*, 88, 333-347.
http://www.stir.ac.uk/staff/psychology/rrm1/STATR7.html