JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

BRUNO LECOUTRE
JACQUES POITEVINEAU
MARIE-PAULE LECOUTRE

Discussion of D. Denis. Fisher : responsible, not guilty

Journal de la société française de statistique, tome 145, n° 4 (2004), p. 55-62.

<http://www.numdam.org/item?id=JSFS_2004__145_4_55_0>

DISCUSSION OF D. DENIS

Fisher: Responsible, not guilty

Bruno LECOUTRE [1], Jacques POITEVINEAU [2],
Marie-Paule LECOUTRE [3]

ABSTRACT

When reading Denis' paper the feeling is that Fisher cannot be judged responsible for the "problems associated with today's model". Even if we agree that current uses of NHST are farm from being pure Fisherian, our analysis is somewhat different. In order to understand the Fisher's real contribution, it is of direct importance to recall his statistical ideas about causality and probability. In particular his works, not only on the *fiducial* theory, but also on the *Bayesian* method in his last years, are a fundamental counterpart to his emphasis on significance tests. In conclusion, while the Fisher's responsibility in the today's practices cannot be discarded, the verdict imposes oneself: "responsible, not guilty"

RÉSUMÉ

La lecture de l'article de Denis donne l'impression que Fisher ne peut pas être jugé responsable des «problèmes associés au modèle d'aujourd'hui». Même si nous sommes d'accord que les usages actuels des tests de signification de l'hypothèse nulle sont loin d'être purement fisheriens, notre analyse est sensiblement différente. Pour comprendre la contribution réelle de Fisher, il est essentiel de rappeler ses idées statistiques sur la causalité et la probabilité. En particulier ses travaux, non seulement sur la théorie *fiduciaire*, mais aussi sur la méthode *bayésienne* dans ses dernières années, constituent une contrepartie fondamentale à son insistance sur l'usage des tests de signification. En conclusion, tandis que la responsabilité de Fisher dans les pratiques actuelles ne peut pas être rejetée, le verdict s'impose de lui même: «responsable, non coupable».

1. ERIS, UMR 6085, Laboratoire de Mathématiques Raphaël Salem C.N.R.S. et Université de Rouen, Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France. E-mail: bruno.lecoutre@univ-rouen.fr
http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm
2. ERIS, UMR 7604, LAM-LCPE, C.N.R.S., Université Paris 6 et Ministère de la Culture, 11 rue de Lourmel, 75015 Paris, France. E-mail: poitevin@ccr.jussieu.fr
3. ERIS, Laboratoire Psy.co, E.A. 1780, Université de Rouen, UFR Psychologie, Sociologie, Sciences de l'Éducation 76821 Mont-Saint-Aignan Cedex, France.
E-mail: marie-paule.lecoutre@univ-rouen.fr

References to the Denis' paper are indicated by DJD

In spite of all the rhetoric that denounced its widespread misinterpretations, null hypothesis significance testing (NHST) still remains the most ubiquitous statistical inference procedure, even when confidence intervals, likelihood, or Bayesian methods are clearly more appropriate. The main goal of Daniel Denis's article is an attempt to demonstrate the little resemblance of today's uses of NHST by social scientists with the original Fisher model. It is also argued that the current model is "hybridised, misused and misunderstood".

The author must be congratulated to have addressed a so controversial domain as the foundations of statistical inference. Understanding the real Fisher's contribution to the current statistical practices is more than a simple historical overview. It is of a great practical importance, and we appreciate the opportunity to comment on it.

1. The Fisher's responsibility for current practices

When reading Denis' paper the feeling is that Fisher cannot be judged responsible for the "problems associated with today's model". Even if we agree that current uses of NHST are far from being pure Fisherian, our analysis is somewhat different. Particularly demonstrative are the two Fisher's papers (1928, 1929) in the *Proceedings of the Society for Psychical Research*, in which he commented on a psychological experiment about card guessing previously published in the same journal. These papers discussed the statistical method in psychical research and were directly addressed to social scientists, which invalidates Denis' statement that "Fisher *never* recommended his procedures for social science" (DJD, section 3) (see also the Fisher's famous "Lady testing tea" example).

Fisher wrote "Personally, the writer prefers to ... *ignore entirely* all results which fail to reach that [significance] level" (Fisher, 1926, page 504, italics added), or again "The test of significance only tells him [the practical investigator] *what to ignore*, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design experiment that will rarely fail to give a significant result" (Fisher, 1929, page 190, italics added). At the very least, it is not an incitement to account for non-significant results, even worse to publish them, and this does not support Denis' claim "Had significance testing remained Fisherian, the file drawer problem would likely not exist today" (DJD, section 1.6). Note that the last sentence of the 1929 citation is quasi identical to the Fisher's citation (1966, page 14) that Denis overuses when assuming it "implied that both significant and non-significant results should be published". Moreover, this shows that the "*original* Fisher model" evolved little.

Furthermore, Fisher explicitly stated "It is a common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary,

but convenient, level of significance for the practical investigator". So, social scientists can hardly be blamed to cite Fisher "as support for their choice of the 0.05 level of significance" (DJD, section 2.3.2), even if Fisher later came to repudiate any systematic predetermined level of significance.

As far as we are concerned with significance testing, the 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 (1990/1925, page 8). This "fundamental paradox", denounced as early as 1942 by Berkson, is a common feature of Fisher's model and today's uses of NHST.

One can hardly understand the Fisher's contribution while saying nothing about his concepts of causality and probability, which are of direct importance for the objectives Fisher assigned to statistical methods.

In particular, in the Fisher's approach, randomization (DJD, section 1.2) and significance tests are strongly linked to causality (see Lecoutre, 2004): "The fact is that if two factors, A and B, are associated – clearly, positively, *with statistical significance* as I say – it may be that A is an important cause of B, it may be that B is an important cause of A, it may be that something else, let us say X, is an important cause of both. If, now, A *the supposed cause has been randomized* – has been randomly assigned to the material from which the reaction is seen – then one may exclude at a blow the possibility that B causes of A, or that X causes A. We know perfectly well what causes A – the fall of the dice or the chances of the random sampling numbers, and nothing else" (Fisher, 1959, page 14, italics added). This has certainly greatly influenced the perception of the role of NHST by social scientists. So a common presentation of this procedure is that rejecting the null hypothesis implies rejecting randomness: for instance, Tryon (2001) wrote "rejection of the null hypothesis implies that the results are not due to chance and that therefore they must be both systematic and reproducible". Moreover, this could explain why "the most commonly occurring weakness in the application of Fisherian methods is, I think, undue emphasis on tests of significance, and failure to recognize that in many types of experimental work, estimates of the treatment effects, together with estimates of the error to which they are subject, are the quantities of primary interest." (Yates 1964; see also Street, 1990).

It is also important to recall that Fisher was evidently interested in *inverse probabilities*, as it emerges from his works not only on the *fiducial* approach but also on the Bayesian method in his last years (Fisher 1962). Added to his firm opposition to the interpretation of the "p-value" as the relative frequency of error when sampling repeatedly in a same population, it is likely to have caused some confusions among scientific workers.

2. The "hybridism" of NHST

This was identified long before Gigerenzer, although this particular term was not used: see for example Morrison and Henkel 1970, page 7. More

than thirty years later the situation has little, if not evolved. Rather than stimulating the interest of experimental scientists, to repeat academic debates and controversies gives a discouraging feeling of déjà-vu. This is without doubt detrimental to the image of statistical inference.

Furthermore, our empirical studies about the way accustomed users – psychological researchers and professional applied statisticians – interpret NHST outcomes revealed us that the attitude of these users was far from being as homogeneous as might be expected (Poitevineau et Lecoutre, 2001; Lecoutre, Lecoutre and Poitevineau, 2003). In fact, it does not exist *a single* hybrid model, but *a variety* of "more or less hybrid" (and in particular more or less Fisherian) attitudes. We agree with Denis that his empirical example about the misuse of significant testing in section 4 reflects a common practice in experimental publication. However, it must be acknowledged that this practice is reinforced by a natural cognitive tendency to "take a position" when being published and in some way to arrange every NHST outcome in a "cognitive filing cabinet", where a significant test goes under "there is an effect" and a nonsignificant test is improperly filed under "there is no effect" (see the significance hypothesis of Oakes, 1986). It is not really a "decision" in the sense of Neyman and Pearson (or of the Bayesian decision-theoretic approach). On the contrary, in our empirical studies, only a minority of accustomed users had a systematically clear-cut attitude. Most users tried to qualify the interpretation of the significance test in relation to the estimate of the treatment effect.

3. Miscellaneous remarks

Fisher always resisted the idea of alternative hypothesis, and his presentations of significance tests were repeatedly in terms of a fundamental "logical distinction": "[in case of significance] *Either* the hypothesis is untrue, *or* the value of χ^2 has attained by chance an exceptionally high value" (Fisher, 1990/1925, page 80). As emphasized by Kruskal (1980, page 1021) about this dichotomy, "one cannot as a rule make sense of the idea without thinking of alternative hypotheses...". While Denis rightly quotes that for Fisher "the null hypothesis ... is possibly disproved" (DJD, section 1.4), he writes unfortunately that "it is questionable whether one can infer it [the alternative hypothesis] when the null is shown to be false" (DJD, section 1.4). If, when the null hypothesis is shown to be false [4], its logical complement (as suggested by Denis) cannot be inferred, what is left? Surely, no statistical test is needed to infer nothing.

Section 2.1 appears to be inconsistent with section 1.3. Actually, it looks paradoxical both to regret the lack of truly random samples (DJD, section 2.1) and to recall the hypothetical character, according to Fisher, of the population (DJD, section 1.3).

4. It is this point that could be questionable: has a significant result really shown that the null is false? That is seriously questioned by some Bayesians, see *e.g.* Berger, 2003.

The *sacredness* of 0.05 is undoubtedly a problem, but is not justified by anyone of the statistical models (Fisherian, Neyman-Pearsonian, Bayesian). Rather, this issue heavily relies on fundamental conceptions that would deserve a much more thorough discussion. Fisher considered the probability that the observed value of the test statistic "will be exceeded by chance" (the "p-value") to characterize a *unique* experiment, and this probability has to be compared with the significance level the researcher has in mind. On the contrary, Neyman-Pearson conceived the experiment only as a member of a set of identical ones and spoke of α, the risk of the first kind, not of significance level.

Concerning the Neyman and Pearson's notion of power (section 1.7), it is rather surprising that the single reference is the 1928 article where the term "power" *does not appear*. Actually "power" really appeared only in their 1933b paper (although, of course, the concept was implicit in the 1933a paper, which establishes the famous "Neyman-Pearson lemma"). In the same section 1.7 it is largely exaggerated to say that "Cohen (1962) later contributed enormously to the concept of power...". We agree that Cohen made a great deal in familiarizing experimentalists with power "by providing relatively easy computational methods", but he did not contribute at all to the *concept* itself. It is also too quickly said, "as has been shown by Cohen, power *does* have a place in scientific experiments". What place is intended should be made explicit. The role of power – in planning of experiments (what sample size?) and/or in interpreting results of experiments – has been a matter of dispute. Nowadays, a more and more widespread opinion is that "for interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead" (Goodman and Berlin, 1994). So, the American Psychological Association recommended: "once the study is analyzed, confidence intervals replace calculated power in describing results" (Wilkinson and Task Force on Statistical Inference, 1999). This agrees with Fisher's position with regards to sensitivity.

The expression "exact level" (DJD, section 1.5) to characterize the "p-value" is unfortunate and misleading, since it is a data-dependent measure. It suggests that the p-values can be interpreted as Type I errors, which is a frequently denounced misinterpretation (by Fisher himself in particular). Moreover, "exact level" has usually a different meaning in statistics (*i.e.* "exact size"). Of course, we can speak about "the exact value of p" (as Fisher did: *e.g.*, Fisher, 1990/1925, page 80), but this does not mean that it can be interpreted as "the exact level of significance". In fact, Fisher (rightly) used the word "exact" when he referred to ("fiducial") probabilities about parameters, given the data in hand, for instance: "the statement can be made that *the probability that the unknown mean* of the population is less than a particular limit, *is exactly* P" (Fisher, 1958, page 271, italics added).

Section 2.3.1 deals with the important issue of the hypotheses to be tested. However, a substantive hypothesis is not "that is held to best account for the data". That one is precisely a statistical hypothesis. The alternative statistical hypothesis may be a point hypothesis (thus as precise as the null

hypothesis), or a composite one (as is typically the case). A statistical test, even the dreadful current hybrid NHST, only deals with statistical hypotheses (although users frequently misuse them in this respect). A substantive hypothesis is the incorporation of a statistical hypothesis (it could be the null hypothesis either!) in a statement within the scientific domain of the study (biology, psychology, physics...). The citation of Fisher (1966) is not relevant here; it is of prime importance, but regarding the fundamental character of the test, *i.e.* whether it is an instrument for increasing knowledge or for action.

Unfortunately, Denis' commendable attempt to contrast the different approaches in Table 1 is unconvincing. This is partly due to the fact that neither NHST users nor Bayesians can be considered an homogeneous class, while on the contrary the distinction between "early Fisher" and "late Fisher" is questionable. As a consequence, most cells appear either as approximations or as out of context claims (especially for the Bayesian column that presents hardly reconcilable personal viewpoints from different authors). This would deserve a lengthy discussion that is beyond the scope of this comment.

Moreover, for the comparison of the Fisherian, Neyman-Pearson and Bayesian approaches, at least three additional papers should be considered. Lehmann (1993) argued "that in their main practical aspects the two theories [Fisher and Neyman-Pearson] are complementary rather than contradictory, and that a unified approach is possible that combines the best feature of both". With an again more ambitious perspective, Berger (2003) discussed the conditional frequentist approach to testing, which is argued "to provide the basis for a methodological unification of the approaches of Fisher, Jeffreys and Neyman". Finally, the famous Savage's (1976) talk "on rereading R.A. Fisher" gave the sympathetic views of a Bayesian on the Fisher's statistical ideas.

4. Conclusion: non guilty

Having recognized the Fisher's responsibility in the today's practices, we are now comfortable to present his defence.

If we restrict the debate to NHST, Perlman and Wu (1999) gave a formal argument which showed that in several composite null hypothesis testing problems optimal tests in the Neyman-Pearson sense are flawed. This is of considerable practical importance in order to avoid unwarranted and inappropriate inference procedures currently in use (Lecoutre, 2005). The authors concluded: "We hope that we have alerted statisticians to the dangers inherent in uncritical application of the NP [Neyman and Pearson] criterion, and, more generally, convinced them to join Fisher, Cox and many others in carefully weighing the scientific relevance and logical consistency of any mathematical criterion proposed for statistical theory" (Perlman and Wu, 1999, page 381).

If we enlarge the debate, it must be stressed that, as a counterpart to his emphasis on significance tests, Fisher was constantly concerned with considering a method that only expressed evidence from data in terms of probability about parameters and had *good conventional* properties. He

considered the *fiducial* approach as ideally suited for this purpose "for there is no other method ordinarily available for making correct statements of probability about the real world" (Fisher, 1935a, pages 198-199). Fiducial inference is admittedly considered by most modern statisticians as a blunder, but it could be speculated with Efron that "maybe Fisher's biggest blunder will become a big hit in the 21st century" (Efron, 1998, page 107). We agree with him that "a widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance" (Efron, 1998, page 106). In actual fact we suggested that "such a theory is by no means a speculative viewpoint but on the contrary a desirable and perfectly feasible project" (Lecoutre, Lecoutre et Poitevineau, 2001). Of course, this is another debate.

In conclusion, the verdict imposes oneself: "responsible, not guilty".

Additional references

BERGER J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? [With discussion]. *Statistical Science, 18*, 1-32.

BERKSON J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

EFRON B. (1998). R.A. Fisher in the 21st century [With discussion]. *Statistical Science, 13*, 95-122.

FISHER R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 503-513.

FISHER R. A. (1928). The Effect of Psychological Card Preferences. *Proceedings of the Society for Psychical Research, 38*, 269-271.

FISHER R. A. (1929). The Statistical Method in Psychical Research. *Proceedings of the Society for Psychical Research, 38*, 189-192.

FISHER R. A. (1958). The Nature of Probability. *Centennial Review, 2*, 261-274.

FISHER R. A. (1959). *Smoking. The cancer controversy*, Edindurgh: Oliver and Boyd.

FISHER R. A. (1962). Some examples of Bayes's method of the experimental determination of probabilities *a priori*. *Journal of the Royal Statistical Society, Series B, 24*, 118-124.

GOODMAN S. N., & BERLIN J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*, 200-206.

KRUSKAL W. H. (1980). The significance of Fisher: A review of R.A. Fisher, the life of a scientist. *Journal of the American Statistical Association, 75*, 1019-1030.

LECOUTRE B. (2004). Expérimentation, inférence statistique et analyse causale. *Intellectica, 38*, 193-245.

LECOUTRE B. (2005). The right use of interval estimates in ANOVA. Submitted for publication.

LECOUTRE M.-P., POITEVINEAU J., & LECOUTRE B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology, 38*, 37-45.

LEHMANN E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association, 88*, 1242-1249.

MORRISON D. E., & HENKEL R. E. (Eds.) (1970). *The Significance Test Controversy – A Reader.* London: Butterwoths.

NEYMAN J., & PEARSON E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231*, 289-337.

NEYMAN J., & PEARSON E. S. (1933b). The testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society, 29*, 492-510.

OAKES M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* New York: Wiley.

PERLMAN M. D., & WU L. (1999). The emperor's new tests. *Statistical Science, 14*, 355-369.

SAVAGE L. J. (1976). On rereading R.A. Fisher [With discussion]. *Annals of Statistics, 4*, 441-500.

STREET D. J. (1990). Fisher's contributions to agricultural statistics. *Biometrics, 46*, 937-945.

TRYON W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological-methods, 6*, 371-386.

WILKINSON L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54*, 594-604.

YATES F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics, 20*, 307-321.