Beyond the significance test controversy: Prime time for Bayes?

Bruno Lecoutre
UPRESA 6085, Analyse et Modèles Stochastiques, C.N.R.S. et Université de Rouen
Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France
E-mail: bruno.lecoutre@univ-rouen.fr

1. Prime time for a widely accepted objective Bayes theory

The experimental research is facing a paradoxical situation. On the one hand, Null Hypothesis Significance Testing (NHST) is required in most publications as an unavoidable norm, but on the other hand, it leads to innumerable misinterpretations and misuses. Moreover, from the outset (Boring, 1919 ; Tyler, 1931; Berkson, 1938; etc.), NHST has been criticized. Its use has been explicitly denounced by the most eminent and the most experienced researchers, both on theoretical and methodological grounds. Sharp controversies have opposed Fisher to Neyman and Pearson on the very foundations of statistical inference. In the sixties, there were more and more criticisms, especially in the behavioural and social sciences, denouncing the shortcomings of NHST and its inadequacy to the purposes of experimental data analysis. Nowadays, while the users' uneasiness is ever growing (Lecoutre M.-P., 1998/1991), academic debates are repeated in these domains and give a discouraging feeling of déjà-vu. To take only an example, the "hybridism" of Fisher and Neyman/Pearson's theories was identified long before Gigerenzer, although this particular term was not used (see, *e.g.*, Morrison & Henkel, 1970, p. 7). Moreover, many recent papers are replete with ill-informed, secondary sources, or ill-considered claims, and at first place concerning Fisherian and Bayesian inferences. Unfortunately, this confusing controversy, rather than stimulating the interest of scientists, continually reinforces their inertia and their resistance to new methods. Nevertheless, it seems to be nowadays a crucial period of time in which changes in reporting experimental results, especially in presenting and interpreting effect sizes, are more and more required by editorial policies (see *e.g.*, Berry, 1986; Braitman, 1988, 1991; Loftus, 1993; Thompson, 1994, 1996; Heldref Foundation, 1997; Murphy, 1997).

So time's up to come to a positive agreement for procedures that bypass the common misuses of NHST, while filling up its role of "an aid to judgement" which "should not be confused with automatic acceptance tests, or 'decision functions'." (Fisher, 1990/1925, p. 128), and satisfying the requirements of scientists for experimental data analysis, in particular the need for *objective* statements and the need for procedures about effect sizes. Undoubtedly, there is increasing acceptance that Bayesian inference is ideally suited for this purpose. Without dismissing the interest of the subjective decision theoretic Bayesian theory, it must be recognized that the Bayesian paradigms also admits objectivity (following Laplace, 1986/1825; see *e.g.*, Jeffreys, 1961 and Jaynes, 1983) and is in this form appropriate for situations involving scientific reporting. More precisely, I suggest that "a *widely accepted* objective Bayes theory, which *fiducial inference* was intended to be, would be of immense theoretical and practical importance. A *successful objective* Bayes theory would have to provide *good frequentist properties in familiar situations*, for instance, reasonable coverage probabilities for whatever replaces confidence intervals." (Efron, 1998, italics added). The purpose of this paper is to argue that a widely accepted Bayes theory is in no way a speculative viewpoint but is a desirable and perfectly feasible project.

2. Its desirability

"*From the table the probability is 0.9985* […] *that 2 is the better soporific*" (Student, 1908). Curiously enough, many critics and defenders of NHST who discuss its foundational aspects ignore Fisher's conception of probability, which is of direct importance for the objectives Fisher assigned

to statistical inference. Fisher firmly argued against the interpretation of the observed level as the relative frequency of error when sampling repeatedly in a same population (Fisher 1990/1956, pp. 81-82). Explicitly, his presentation of Student's *t* test did not refer to a frequentist conception (conditional on parameters), but on the contrary involved a predictive distribution conditional to the observed standard deviation (Lecoutre, 1985). Like Bayesians, Fisher was evidently interested in inverse probability, as evidenced not only by his work on the *fiducial* theory (*e.g.*, Fisher 1990/1956), but also on the Bayesian method in his last years (Fisher, 1962). He had the constant concern for considering a method that expresses only evidence from the data in terms of probability about parameters and has good conventional properties, which are the necessary condition for a *widely accepted objective* theory. Fiducial inference is admittedly considered by most modern statisticians to be a blunder, but it could be speculated with Efron (1998) that "maybe Fisher's biggest blunder will become a big hit in the 21st century".

"*It would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms*" (Rouanet, 1998, p. 54). A more and more widespread opinion among applied statisticians 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 & Berlin, 1994). All these methods are intended to deal with the question of *effect sizes*, which is essential "because Science is inevitably about magnitudes" (Cohen, 1990). They can at least prevent the two main erroneous interpretations of NHST which consists, on the one hand in confusing statistical significance with substantive significance (one of the most often denounced error: *e.g.*, Selvin, 1957; Kish, 1959; Bolles, 1962; Bakan, 1966; *etc.*), and on the other hand in interpreting a nonsignificant result as proof of the null hypothesis (an error which can be found in many experimental publications, even in prestigious journals, as noted by Harcum, 1990). At the present time, the *official* trend for experimental publications is to advocate the use of confidence intervals, in addition or in place of NHST (see, *e.g.*, American Psychological Association, 1996). Unfortunately, it is so strange to treat the data as random even after observation that the frequentist interpretation of confidence intervals does not make sense for most of the users. Ironically, it is their "incorrect natural" interpretation in terms of (Bayesian) probabilities about parameters which is their appealing feature. Moreover the success of significance tests and confidence intervals is built on the *duplicity* of most statistical instructors, who tolerate these heretic interpretations, and even often use them. Thus it can be anticipated that the conceptual difficulties encountered with the frequentist conception of confidence intervals will produce further dissatisfaction.

"*Why are experimental psychologists (and others) reluctant to use Bayesian inferential procedures in practice?*" (Winkler, 1974). In a very lucid paper, which seems have been written today, Winkler answered that "this state of affairs appears to be due to a combination of factors including philosophical conviction, tradition, statistical training, lack of 'availability', computational difficulties, reporting difficulties, and perceived resistance by journal editors". If we leave on one side the choice of a philosophical approach to statistical inference which is "not really as important as whether the approach is used consistently, carefully, and appropriately", none of these arguments has sound ground. However, Bayesian methods often encounter the mistrust, if not the automatic opposition, of scientists who felt that they were too complicated to use and too subjective to be scientifically acceptable. The recent comment by Falk and Greenbaum (1995) that "Bayesian inference might, in principle, fill the void created by abandoning significance-testing", but that "implementation of Bayesian analysis, however, requires subjective assessments of prior distributions, and often involves technical problems" illustrates this attitude. Also the contribution of Bayesian inference to experimental data analysis has often been obscured by the insistence of many authors for pointing out the merits of the Bayesian approach in decision making. "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested." (Rozeboom, 1960). In consequence, without speaking of irrelevant caricature-like considerations (*e.g.*, Chow, 1996), Bayesian methods for analysing experimental data have been constantly ignored or discarded (*e.g.*, Wilson, Miller & Lower, 1967; Frick, 1996) for a priori

reasons that are more and more unjustified. Moreover, the dominant frequentist conception, and the widespread use of significance tests, still appear to be such a steamroller (as says Berry, 1993) that even those who are sympathetic towards the Bayesian approach often drop the Bayesian label to get easier acceptance of their proposals. For instance, in a methodological paper for medical researchers, Goodman and Berlin (1994) give a very persuasive preliminary presentation of Bayesian methods. But, after having declared that "Bayesian posterior probabilities are exactly what scientists want", they only discuss the use of confidence intervals, arguing that they are "more familiar" to readers than Bayesian probabilities.

"*At the very least, use of noninformative priors should be recognized as being at least as objective as any other statistical techniques*." (Berger, 1985, p. 110). It must be acknowledged that any widely accepted inferential methods require some more or less arbitrary choices and conventions. So the arbitrariness of the choice of α has been pointed out for a long time (*e.g.*, Rozeboom, 1960; Camilleri, 1962; Winer, 1962, p. 13; *etc.*). J. Neyman himself recognized an element of subjectivity in the theory of tests he founded with E. Pearson, for he firmly stated that the hypothesis to be tested (the so-called "null hypothesis", though not in Neyman's words) should be the one for which the risk of rejecting it if true is the most important to control and this, he admitted, is a subjective matter (Neyman, 1950). The noninformative Bayesian approach, based on vague priors, cannot avoid conventions either. For instance, in a Bernouilli process, the NSHT procedure involves arbitrariness both in the specification of a stopping rule (Lindley & Phillips, 1976) and in the choice of whether or not include the observed data's probability in the *p* value. This arbitrariness within the frequentist approach have an exact counterpart in the particular choice of a Bayesian prior distribution in an "ignorance zone" (Bernard, 1996). But, based on more useful working definitions than frequentist procedures, Bayesian procedures make all the choices explicit and then more easily questionable. Moreover it offers considerably more flexibility and provides relevant answers to virtually all the questions asked by experimental data analysis.

3. Its feasibility

For many years, with other colleagues in France, we have worked in a fiducial (for motivation) and Bayesian (for technique) perspective in order to develop standard ("noninformative") Bayesian methods for most *familiar situations* encountered in experimental data analysis. Our conclusion is that these methods are *nowadays* available and can be easily taught and used. They are concrete proposals for bypassing the shortcomings of NHST and improving the current statistical methodology and practice (Rouanet *et al.*, 1998/1991). Our statistical consulting experience, especially in psychology, revealed us that they were far more intuitive and much closer to the thinking of scientists than frequentist methods (see also Kadane, 1995). They have been applied many times to real data and well-accepted by psychological journals (see *e.g.*, Hoc & Leplat, 1983; Ciancia *et al.*, 1988; Lecoutre, 1992; Hoc, 1996; Clément & Richard, 1997; and many experimental articles published in French).

Standard (or fiducial) methods: Beyond significance tests. A well-known feature of Bayesian inference is that it can be used to reinterpret many of the frequentist procedures. For instance, for the comparison of two means with the usual *t* test, the one-sided *p*-value is exactly the standard Bayesian probability that the true difference has the opposite sign of the observed difference. This is precisely Student's highly meaningful interpretation that "the probability is 0.9985 [1-*p*] that 2 is the better soporific". From this interpretation, it becomes straightforward to effectively fight the erroneous interpretations of NHST. Moreover, the possibility of displaying and interactively investigating posterior distributions by means of visual software nowadays gives an attractive conceptual simplicity to Bayesian procedures and allows the users to easily understand their many appealing features. In the first place, a decisive contribution is that procedures for assessing the magnitude of effects are immediately available (the reader is more especially referred to Rouanet & Lecoutre, 1983; Lecoutre, Derzko & Grouin, 1995; Lecoutre, 1996; Rouanet, 1996; Bernard, 1998/1991; Lecoutre & Derzko, 1999).

Other Bayesian techniques are promising. Standard Bayesian methods undoubtedly have a privileged status in order to get "public use" statements for reporting results that incorporate and extend significance tests. But other Bayesian techniques, based on "informative" prior distributions, have presumably also an important role to play in experimental investigations. They are ideally suited for making "personal" decisions and for integrating multiple studies into meta-analyses. Realistic uses of these techniques have been proposed. On the one hand, various prior distributions expressing results from other experiments or subjective opinions of well-informed specific individuals, whether *sceptical* or *enthusiastic,* can be investigated to assess the robustness of the conclusions (*e.g.*, Spiegelhalter, Freedman & Parmar, 1994). On the other hand, a major strength of the Bayesian approach, which is the ease of making predictions concerning events of interest, give efficient tools for designing ("how many subjects?") and monitoring ("when to stop?") experiments. Bayesian predictive probabilities enable the researcher to evaluate the chances that the experiment will end up showing a conclusive, or on the contrary a non-conclusive, result, on the basis either of a preliminary study or of the partial data of a current experiment (*e.g.*, Choi & Pepple, 1989; Berry, 1991; Lecoutre, Derzko & Grouin, 1995).

4. Conclusion

"*Null-hypothesis tests are not completely stupid, but Bayesian statistics are better*." (Rindskopf, 1998). Bayesian routine procedures for the *familiar situations* of experimental data analysis are nowadays easy to implement and to teach. They offer promising *new ways* in statistical methodology. Now our teaching experience is firmly established: using the noninformative Bayesian interpretations of significance tests and confidence intervals in the natural language of probabilities about unknown effects comes quite naturally to students. In return their common misuses appear to be more clearly understood. So, rather than banning NHST, which seems to be a highly unrealistic device, a more salutary task would be to promote its Bayesian interpretation.

"*We need statistical thinking, not rituals*" (Gigerenzer, 1998). The Bayesian philosophy, which emphasizes the need to think hard about the real information provided by the data in hand ("what have the data to say?") instead of applying ready-made procedures, should become an attractive challenge for the scientists, the applied statisticians and the statistical instructors of the 21st century.

REFERENCES

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (1998/1991). New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference (1st edition in french entitled "L'Inférence Statistique dans la Démarche du Chercheur", 1991). Peter Lang. Bern.

The other references can be found in a detailed bibliography prepared by B. Lecoutre and J. Poitevineau (available on the Internet at address: http://epeire.univ-rouen.fr/labos/eris/pac.html, or send e-mail to: *bruno.lecoutre@univ-rouen.fr*).

RÉSUMÉ

Cet article défend la thèse que des méthodes bayésiennes objectives et largement admises, répondant à la motivation fiduciaire de Fisher, sont non seulement désirables mais aussi faisables. Ces méthodes permettent de dépasser la controverse sur les tests de signification et ouvrent la voie à une nouvelle méthodologie de l'analyse des données expérimentales.