TEACHING BAYESIAN METHODS FOR EXPERIMENTAL DATA ANALYSIS

Bruno Lecoutre, C.N.R.S. et Université de Rouen Mathématiques, France

The innumerable articles denouncing the deficiencies of significance testing urge us to reform the teaching of statistical inference for experimental data analysis. Bayesian methods are a promising alternative. However, teaching the Bayesian approach should not introduce an abrupt changeover from the current frequentist procedures: at the very least, the two approaches should co-exist for many years to come. According to this fact, we have developed statistical computer programs, that incorporate both current practices and standard Bayesian procedures. These programs are used in the graduate statistics course in psychology, where Bayesian methods are especially introduced for inferences about effect sizes in the analysis of variance framework. Most of them are available on the Internet at address: http://epeire.univ-rouen.fr/labos/eris/pac.html.

THE SHORTCOMINGS OF USUAL SIGNIFICANCE TESTING

Many recent papers have stressed on the necessity of changes in reporting experimental results. This has been recently made official by the American Psychological Association (APA, 1996). A more and more widespread opinion is that procedures that provide genuine information about the magnitude of effects must be used in addition to null hypothesis significance tests. Such procedures have been developed both in the frequentist and Bayesian frameworks. But they are again rarely used, in spite of the fact that they are nowadays straightforward to implement. So it must be urged to reform the teaching of statistical inference and to include these procedures, even in introductory courses.

THE DUPLICITY OF STATISTICAL INSTRUCTORS

At the present time, the *official* trend advocates the use of confidence intervals (*e.g.*, APA, 1996). But it can be anticipated that the conceptual difficulties encountered with the frequentist conception of confidence intervals will produce further dissatisfaction. In particular the users will realize that the appealing feature of confidence intervals is the result of a fundamental misunderstanding. In the frequentist framework, *one* 95% confidence interval cannot be thought of as a *fixed* interval having a 95% chance of including the true value of interest. The “correct” frequentist interpretation of confidence intervals, as for the significance tests, involves a long run repetition of the same experiment and does not make sense for most of the users. It is the interpretation of confidence intervals in terms of probabilities about parameters that is their appealing

feature: “Again it is not clear why such a set should be of interest unless one makes the *natural error* of thinking of the parameter as random and the confidence set as containing the parameter with a specified probability.” (Kadane, 1995).

Moreover the success of significance tests and confidence intervals is built on the duplicity of most statistical instructors, who tolerate “incorrect natural” interpretations, and even often use them. This fact is apparent in many statistical articles or textbooks: “...But it is perhaps of greater interest to be able to say whether or not some observed association in a *sample* of scores indicates that the variables under study are most probably associated in the *population* from which the sample was drawn.” (Siegel 1956, page 195); “…a random sample can be used to specify a segment or interval on the number line such that the parameter has a high probability of lying on the segment. The segment is called a confidence interval.” (Kirk 1982, page 42); *etc.* We completely agree with the statement of Freeman (1993) that in attempts to teach the “correct” interpretation of frequentist procedures “we are fighting a losing battle”.

THE BAYESIAN CHOICE

At the very least, it must be recognized that the Bayesian interpretation is far more intuitive and much closer to the thinking of scientists: “Bayesian posterior probabilities are exactly what scientists want” (Goodman and Berlin, 1994). This is clearly an attractive feature in teaching inference (Albert, 1995). But until now, Bayesian methods have often encountered 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 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.

Clearly, teaching Bayesian methods for experimental data analysis should not introduce an abrupt changeover from the current frequentist practices. Given the widespread use of significance tests, this would be highly unrealistic: as Berry (1993) says, “the steamroller of frequentism is not slowed by words.” As a consequence, rather than replacing these practices, Bayesian procedures for experimental data analysis should incorporate, extend, and refine them.

For teaching purposes in the context of experimental data analysis, the "noninformative" Bayesian methods have clearly a privileged status. Based on more useful working definitions than frequentist procedures, they are fully justified, at least as objective, and they can be used as easily as the *t*, *F* or chi-square tests. Moreover, a well-known feature of this standard Bayesian inference is that it provides insightful interpretations of many frequentist procedures. For instance, for the comparison of two means from independent groups with the usual Normal model, which assumes variance equality, the observed one-sided significance level in Student's *t* test can be interpreted as the Bayesian probability that the true difference and the observed difference have opposite signs. Furthermore in this case, the Bayesian credibility interval is identical to the frequentist confidence interval. These interpretations bridge the conceptual and technical gap between Bayesian inference and frequentist procedures, and offer the students a smooth transition from the traditional techniques to the Bayesian methods. Furthermore, the Bayesian interpretations clearly point out the methodological shortcomings of usual null hypothesis significance testing: it is quite apparent from the above example that the significance level only makes a statement about the sign, and has nothing to say about the real size of the difference. On the contrary, Bayesian procedures are ideally suited to drawing conclusions about the magnitude of the investigated effects in a very direct and natural way (see Rouanet, 1996).

Bayesian methods have many other attractive features. In addition to the necessary objective statements for reporting results based on standard Bayesian procedures, they provide efficient tools for personal decisions and for designing ("How many subjects?") and monitoring ("When to stop?") experiments. On the one hand various prior distributions expressing results from other experiments or subjective opinions of well-informed specific individuals, whether *skeptical* or *enthusiastic*, can be investigated to assess the robustness of the conclusions (see *e.g.*, the Bayesian methodology for clinical trials exposed by Spiegelhalter *et al.*, 1994). On the other hand, Bayesian predictive probabilities can be used in a natural way for choosing a sample size and for conducting interim analyses. They enable the scientist to evaluate the real chances of a given conclusion to be obtained with possible future observations, on the basis either of a "pilot" study or of partial results of a current experiment (see *e.g.*, Lecoutre *et al.*, 1995).

THE SPECIFIC INFERENCE APPROACH AND COMPUTER SOFTWARE FOR TEACHING BAYESIAN METHODS

Our work on analysis of variance shows that standard Bayesian procedures can be taught as easily as the traditional *F* ratios. For complex experimental designs, the construction of these procedures is based on the *specific inference* principle (see Rouanet and Lecoutre, 1983; Lecoutre, 1996). In short, this principle consists of considering the effects of interest separately, and making each inference from specifically relevant derived data. This conception brings a simple way for teaching the analysis of variance methods and their Bayesian extensions to non-statisticians (Lecoutre, 1998).

Statistical computer programs have been developed (Lecoutre and Poitevineau, 1992; Lecoutre, 1996). According to our conception, they incorporate both current practices (significance tests, confidence intervals) and standard (noninformative) Bayesian procedures. Prior conjugate distributions are also available. A “Bayesian module” displays and prints Bayesian probability distributions and calculates the corresponding probability statements, in interaction with the user. All of the procedures are applicable to general experimental designs (in particular, repeated measures designs), balanced or not balanced, with univariate or multivariate data, and covariables. These programs are used in the graduate statistics course in psychology, where Bayesian methods are especially introduced in the analysis of variance framework. The possibility of teaching these methods in the context of realistic complex experimental designs, such as repeated measures designs, which are frequently used in experimental research (especially in behavioral sciences), is a decisive advantage for motivating students.

Leaning on these programs, a limited set of theoretic notions is needed to introduce basic procedures, *i.e.* inferences about one degree of freedom effects in complex designs. An introductory course about descriptive statistics, and elementary inference techniques for the comparison of two means, is generally a sufficient background. Then the attention can be concentrated about the interpretations and the practical meaning of procedures. As a consequence, the principles of advanced techniques can be more easily understood, independently of their mathematical difficulty.

Our teaching experience is now firmly established. On the one hand, using the Bayesian interpretations of significance tests and confidence intervals, on the basis of the standard posterior distributions, comes quite naturally to students. In return the frequentist approach and its methodological shortcomings, when restricted to usual significance

testing, appear to be more clearly understood. In a certain sense, within the Bayesian approach, once the posterior distribution has been obtained, only one procedure is involved: given the conclusion searching for (for instance a small or on the contrary a large true effect), the probability of a relevant statement is computed. This conceptual simplicity of the Bayesian approach is here a decisive advantage in teaching procedures for assessing size of effects.

On the other hand, the mechanics of Bayesian inference can be learned by interactively investigating prior distributions. This allows the students to understand the relative roles of sample sizes, data and external information. Predictive distributions appear also as a natural tool. They are especially introduced for choosing appropriate sample sizes. The notion of power generally used for this purpose is introduced here as a limiting case.

REFERENCES

Albert, J. (1995). Teaching inference about proportions using Bayes and discrete models. *Journal of Statistics Education*, *3*(3). Available e-mail: archive@jse.stat.ncsu.edu, message: send jse/v3n3/albert.

APA (1996). American Psychological Association, Board of Scientific Affairs, Task force on statistical inference initial report (draft). Internet: http://www.apa.org/science /tfsi.html.

Berry, D. A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine*, *12*, 1377-1393.

Falk, R., and Greenbaum, C. W. (1995). Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*, 75-98.

Freeman, P. R. (1993) – The role of *p*-values in analysing trial results. *Statistics in Medicine*, *12*, 1443-1452.

Goodman, S. N., and 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.

Kadane, J. B. (1995). Prime time for Bayes. *Controlled Clinical trials*, *16*, 313-318.

Kirk, R. E. (1982). *Experimental Design*. Belmont: Brook-Cole. (2nd edition.)

Lecoutre, B. (1996). *Traitement Statistique des données expérimentales: Des pratiques traditionnelles aux pratiques bayésiennes - Avec programmes Windows par B. Lecoutre et J. Poitevineau*. Saint-Mandé: C.I.S.I.A.

Lecoutre, B. (1998). Teaching analysis of variance and procedures for assessing the magnitude of effects: The specific analysis approach. Submitted for publication.

Lecoutre, B., Derzko, G., and Grouin, J-M. (1995). Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, *14*, 1057-1063.

Lecoutre, B., and Poitevineau, J. (1992). PAC (*Programme d'Analyse des Comparaisons*) - *Guide d'utilisation et manuel de référence*. Saint-Mandé (France): C.I.S.I.A..

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, *119*, 149-158.

Rouanet, H., and Lecoutre, B. (1983). Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, *36*, 252-268.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian Approaches to randomized trials. *Journal of the Royal Statistical Society A, 157*, 357-416.