Journal of Data Science **4**(2006), 207-232

Training Students and Researchers in Bayesian Methods

Bruno Lecoutre
Université de Rouen

Abstract: Frequentist Null Hypothesis Significance Testing (NHST) is so an integral part of scientists' behavior that its uses cannot be discontinued by flinging it out of the window. Faced with this situation, the suggested strategy for training students and researchers in statistical inference methods for experimental data analysis involves a smooth transition towards the Bayesian paradigm. Its general outlines are as follows. (1) To present natural Bayesian interpretations of NHST outcomes to draw attention to their shortcomings. (2) To create as a result of this the need for a change of emphasis in the presentation and interpretation of results. (3) Finally to equip users with a real possibility of thinking sensibly about statistical inference problems and behaving in a more reasonable manner. The conclusion is that teaching the Bayesian approach in the context of experimental data analysis appears both *desirable* and *feasible*. This feasibility is illustrated for analysis of variance methods.

Key words: ANOVA, experimental data analysis, interval estimates, magnitude of effects, p-values, teaching Bayesian methods.

1. Introduction

Today is a crucial time because we are in the process of defining new publication norms for experimental research. In psychology the necessity of changes in reporting experimental results has been recently made official by the American Psychological Association (Wilkinson *et al.*, 1999; American Psychological Association, 2001). In all experimental fields, and especially in medical research, this necessity is supported more and more by journal editors who require authors to routinely report effect size indicators and their interval estimates, in addition to or in place of the results of traditional Null Hypothesis Significance Testing (NHST).

The present paper is divided into four sections. (1) I argue that NHST is an inadequate method for experimental data analysis, not because it is an incorrect normative model, just because it does not address the questions that scientific research requires. I present and criticize the recommendations proposed by the

Task Force of the American Psychological Association to overcome this inadequacy. (2) As an alternative, I suggest teaching Bayesian methods as a *therapy* against the misuses and abuses of NHST. (3) The feasibility of this teaching is illustrated in the context of analysis of variance methods. (4) Its advantages and difficulties are discussed. In conclusion, training students and researchers in Bayesian methods should become an attractive challenge for statistical instructors.

2. The Current Context of the Experimental Research

2.1 The stranglehold of null hypothesis significance tests

From the outset (Boring, 1919; Tyler, 1931; Berkson, 1938; etc.), NHST has been subject to intense criticism, both on theoretical and methodological grounds, not to mention the sharp controversy that opposed Fisher to Neyman and Pearson on the very foundations of statistical inference. In the sixties there was more and more criticism, especially in the behavioral and social sciences (see especially Morrison and Henkel, 1970). The fundamental inadequacy of NHST in experimental data analysis has been denounced by the most eminent and most experienced scientists (see Poitevineau, 1998; Lecoutre, Lecoutre and Poitevineau, 2001).

Several empirical studies emphasized the widespread existence of common misinterpretations of NHST among students and psychological researchers (Rosenthal and Gaito, 1963; Nelson, Rosenthal and Rosnow, 1986; Oakes, 1986; Zuckerman, Hodgins, Zuckerman and Rosenthal, 1993; Falk and Greenbaum, 1995; Mittag and Thompson, 2000; Gordon, 2001; Poitevineau and Lecoutre, 2001). Recently, Haller and Krauss (2002)[1] found out that most methodology instructors who teach statistics to psychology students, including professors who work in the area of statistics, share their students' misinterpretations. Furthermore, Lecoutre, Poitevineau and Lecoutre (2003) showed that professional applied statisticians from pharmaceutical companies are not immune to misinterpretations of NHST, especially if the test is nonsignificant.

If some of the above results could be interpreted as an individual's lack of mastery, this explanation is hardly applicable to professional statisticians. More likely these results reveal that NHST does not address the questions that scientific research requires. Thus, users must resort to a more or less "naïve" mixture of NHST results and other information. In other words they must make "judgmental adjustments" (Bakan, 1966; Phillips, 1973, p.334) or "adaptative distorsions" (M.-P. Lecoutre, 2000, p.74) designed to make an ill-suited tool fit their true

[1]Retrieved March 9, 2006 from http://www.mpr-online.de.

needs. So the confusion between *statistical* significance and *scientific* significance ("the more significant a result is, the more scientifically interesting it is, and/or the larger the true effect is") illustrates such an adjustment and can be seen as an *adaptative abuse*. The improper uses of nonsignificant results as "proof of the null hypothesis" is again more illustrative; indeed, faced with a nonsignificant result, users seem to have no other choice but to either interpret it as proof of the null hypothesis or attempt to justify it by citing an anomaly in the experimental conditions or in the sample. Also the "incorrect" interpretations of p-values as "inverse" probabilities (1-p is "the probability that the alternative hypothesis is true" or is considered as "evidence of the replicability of the result"), even by experienced users, reveal questions that are of primary interest for the users. Such interpretations suggest that "users really want to make a different kind of inference" (Robinson and Wainer, 2002, p.270). Moreover, many psychology researchers explicitly state that they are dissatisfied with current practices and appear to have a real consciousness of the stranglehold of NHST (M.-P. Lecoutre, 2000). They use significance tests only because they know no other alternative, but they express the need for inferential methods that would be better suited for answering their specific questions. In this context a consensus consists in expecting the statistical analysis to express in an objective way "what the data have to say" independently of any outside information. Indeed very few researchers state that they want to integrate outside information – notably theoretical background – into the statistical analysis of data.

2.2 Time for change in teaching statistical inference methods

These findings encourage the many recent attempts to improve the habitual ways of analyzing and reporting experimental data. We can expect with Kirk (2001, p.217) that these attempts "will set off a chain reaction" and in particular that "teachers of statistics, methodology, and measurement courses will change their courses" and that "faculties will require students to learn the full arsenal of quantitative and qualitative statistical tools". We cannot accept that future statistical inference methods users will continue using non appropriate procedures "because they know no other alternative".

So the time has come to create a shift of emphasis in the teaching of statistical inference methods, even in introductory courses for non-statistician students. A more and more widespread opinion is that inferential procedures that bypass the common misuses of significance tests while providing genuine information about the size of effects must be taught *in addition to* (or even *instead of*) NHST. For this purpose, confidence intervals, likelihood, or Bayesian methods are clearly appropriate (e.g., Goodman and Berlin, 1994; Nester, 1996; Rouanet, 1996). Today, the majority trend is to advocate the use of confidence intervals. The following

extracts are proposed guidelines by the Task Force of the American Psychological Association (Wilkinson *et al.*, 1999) for revising the statistical section of the American Psychological Association Publication Manual (italics are mines).

Hypothesis tests. "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval. *Never use the unfortunate expression 'accept the null hypothesis.' Always provide some effect-size estimate when reporting a p value*."

Interval estimates. "*Interval estimates should be given for any effect sizes involving principal outcomes*. Provide intervals for correlations and other coefficients of association or variation whenever possible."

Effect sizes. *Always present effect sizes for primary outcomes*. If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure."

Power and sample size. "Provide information on sample size and the process that led to sample size decisions. *Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations.* Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size."

2.3 Further difficulties

"*It would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms*" (Rouanet, 2000, in Rouanet *et al.*, page 54).

Confidence intervals could quickly become a compulsory norm in experimental publications. However, for many reasons due to their *frequentist* conception, confidence intervals can hardly be viewed as the ultimate method. Indeed the appealing feature of confidence intervals is the result of a fundamental misunderstanding. As is the case with significance tests, the frequentist interpretation of a 95% confidence interval involves a long run repetition of the same experiment: in the long run 95% of computed confidence intervals will contain the "true value" of the parameter; each interval in isolation has either a 0 or 100% probability of containing it. It is so strange to treat the data as random even *after observation* that the *orthodox* frequentist interpretation of confidence intervals does not make

sense for most users. It is undoubtedly the natural (Bayesian) interpretation of confidence intervals in terms of "a fixed interval having a 95% chance of including the true value of interest" which is their appealing feature.

Even experts in statistics are not immune from *conceptual* confusions about frequentist confidence intervals. So, for instance, Rosnow and Rosenthal (1996, p.336) take the example of an observed difference between two means $d = +0.266$. They consider the interval $[0, +532]$ whose bounds are the "null hypothesis" (0) and what they call the "counternul value" ($2d = +0.532$), computed as the symmetrical value of 0 with regard to d. They interpret this specific interval $[0, +532]$ as "a 77% confidence interval" ($0.77 = 1 - 2 \times 0.115$, where 0.115 is the one-sided p-value for the usual t test). If we repeat the experience, the counternull value and the p-value will be different, and, in a long run repetition, the proportion of null-counternull intervals that contain the true value of the difference δ will not be 77%. Clearly, 0.77 is here a data dependent probability, which needs a Bayesian approach to be correctly interpreted.

Beyond these difficulties with frequentist confidence intervals, the proposed guidelines are both partially technically redundant and conceptually incoherent. Just as NSHT, they should result in teaching a set of recipes and rituals (power computations, p-values, confidence intervals...), without supplying a real statistical thinking. In particular, one can be afraid that students (and their teachers) continue to focus on the statistical significance of the result (only wondering whether the confidence interval includes the null hypothesis value) rather than on the full implications of confidence intervals. As the authors of these guidelines state, it is probably true that "*statistical methods should guide and discipline our thinking but should not determine it*." However it is no less true that it would be "folly of blindly adhering to a ritualized procedure" (Kirk, 2001, p.207).

3. The Bayesian Alternative

We then naturally have to ask ourselves whether the "Bayesian Choice" will not, sooner or later, be unavoidable (Lecoutre, Lecoutre and Poitevineau, 2001).

3.1 What is Bayesian inference for experimental data analysis?

"*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).

For the statistician, the role of probabilities, an thus the debates between "frequentists" and "Bayesians", can be expressed in these terms (Lindley, 1993): "whether the probabilities should only refer to data and be based on frequency or whether they should *also* apply to hypotheses and be regarded as measures of

beliefs" (italics added). Bayesian inference, based on a more general and more useful working definition of probability, can address directly problems that the frequentist approach can only address indirectly by resorting to arbitrary tricks.

The most common criticism of the Bayesian approach by frequentists is the need for prior probabilities. Many Bayesians place emphasis on a subjective perspective. An extremist view is that of Savage (1954) who claimed his intention to incorporate prior *opinions* – not only prior knowledge – into scientific inference. Moreover, by their insistence on the decision-theoretic elements of the Bayesian approach, many authors have obscured the contribution of Bayesian inference to experimental data analysis and scientific reporting. This can be the reasons why until now scientists have been reluctant to use Bayesian inferential procedures in practice for analysing their data.

Without dismissing the merits of the decision-theoretic viewpoint, it must be recognized that there is another approach which is just as Bayesian which was developed by Jeffreys in the thirties (Jeffreys, 1998/1939). Following the lead of Laplace (1986/1825), this approach aimed at assigning the prior probability when "nothing" was known about the value of the parameter. In practice, these *noninformative* prior probabilities are vague distributions which, *a priori*, do not favor any particular value. Consequently they let the data "speak for themselves" (Box and Tiao, 1973, p.2). In this form the Bayesian paradigm provides, if not objective methods, at least *reference* methods appropriate for situations involving scientific reporting. This approach of Bayesian inference is now recognized like a standard: "We should indeed argue that noninformative prior Bayesian analysis is the single most powerful method of statistical analysis" (Berger, 1985, p.90).

3.2 Routine Bayesian methods for experimental data analysis

For more than twenty-five years now, with other colleagues in France I have worked in order to develop routine Bayesian methods for the most familiar situations encountered in experimental data analysis (see e.g., Rouanet and Lecoutre, 1983; Lecoutre, Derzko and Grouin, 1995; Lecoutre, 1996; Lecoutre and Charron, 2000; Lecoutre and Poitevineau, 2000; Lecoutre and Derzko, 2001). These methods can be used and taught as easily as the t, F or *chi-square* tests. We argued that they offer promising new ways in statistical methodology (Rouanet *et al.*, 2000).

We have especially developed "noninformative methods". In order to promote them, it seemed important to us to give them a more explicit name than "standard", "noninformative" or "reference". We proposed to call them *fiducial Bayesian* (B. Lecoutre, 2000). This deliberately provocative name pays tribute to Fisher's work on scientific inference for research workers (Fisher, 1990/1925). It indicates their specificity and their aim to let the statistical analysis express

what the data have to say independently of any outside information. Fiducial Bayesian methods are concrete proposals in order to bypass the inadequacy of NHST. They have been applied many times to real data and have been accepted well by experimental journals (see e.g., Hoc and Leplat, 1983; Ciancia *et al.*, 1988; Lecoutre, 1992; Desperati and Stucchi, 1995; Hoc, 1996; Amorim and Stucchi, 1997; Amorim *et al.*, 1997; Clment and Richard, 1997; Amorim *et al.*, 1998; Amorim *et al.*, 2000; Lecoutre *et al.*, 2003, 2004; and many experimental articles published in French).

3.3 The desirability of Bayesian methods

Clearly, the Bayesian approach offers more flexibility to experimental data analysis. In order to illustrate its advantages, I will consider the pharmaceutical example used by Student (1908) in his original article on the t test. Given, for each of the n=10 patients the two "additional hour's sleep" gained by the use of two soporifics [1 and 2], Student used his t test for an inference about the difference of means between the two soporifics, "by making a new series, subtracting 1 from 2" (the ten individual differences are given in Table 1). Then he computed the mean +1.58 [d] and the (uncorrected) standard deviation 1.17 [hence $s = 1.23$, corrected for df] of this series, and concluded from his table of the "t distribution" that "the probability is .9985 or the odds are about 666 to 1 than 2 is the better soporific" (which is not an *orthodox* frequentist formulation!). In modern statements, we compute the t test statistic for the inference about a normal mean t=+1.58/(1.23/$\sqrt{10}$)=+4.06 and we find the one-sided *p-value* 0.0014 (9 *df*).

Table 1: Rection time experiment: Basic data and relevant data for interaction and for group cpmparisons.

+1.2	+2.4	+1.3	+1.3	0	+1.0	+1.8	+0.8	+4.6	+1.4

Some features, outlined hereafter, illustrate the desirability of Bayesian methods that are an alternative to the Task Force Guidelines.

Hypothesis tests: Fiducial Bayesian interpretation of p-values. Fiducial-Bayesian inference provides insightful interpretations of frequentist procedures in intuitively appealing and readily interpretable forms using the natural language of Bayesian probability. For instance, the one-sided p-value of the t test is exactly the fiducial Bayesian probability that the true difference δ has the opposite sign of the observed difference. Given the Student's data ($p = 0.0014$, one-sided), there is a 0.14% posterior probability of a negative difference and a 99.86% complementary probability of a positive difference. In the Bayesian framework these statements are *statistically correct*.

Moreover the fiducial Bayesian interpretation of p-values clearly points out the methodological shortcomings of NHST. It becomes apparent that the p-value *in itself* says nothing about the magnitude of δ. On the one hand, even a "highly significant" outcome (p "very small") only establishes that δ has the same sign as the observed difference d. On the other hand, a "nonsignificant" outcome is hardly worth anything, as exemplified by the fiducial Bayesian interpretation $Pr(\delta < 0) = Pr(\delta > 0) = 1/2$ of a "perfectly nonsignificant" test (i.e. $d = 0$).

Interval estimates: Fiducial Bayesian interpretation of the usual CI. Another important feature is the interpretation of the usual confidence interval in natural terms. In the Bayesian framework, this interval is usually termed a *credibility interval* or a *credible interval*, which explicitly accounts for the difference in interpretation. It becomes correct to say that "there is a 95% probability (or *guarantee*) of δ being included between the fixed bounds of the interval" (conditionally on the data), i.e. for the Student's example between +0.70 and +2.46 hours.

Effect sizes: Straight Bayesian answers. Beyond the reinterpretations of the usual frequentist procedures, other Bayesian statements give straight answers to the question of effect sizes. We can compute the probability that δ exceeds a fixed, easier to interpret, additional time; for instance "there is a 91.5% probability of δ exceeding one hour". Since the units of measurement are meaningful, it is easy to assess the practical significance of the magnitude of δ. To summarize the results, it can be reported that "there is a 91.5% posterior probability of a large positive difference ($\delta > +1$), a 8.4% probability of a positive but limited difference ($0 < \delta < +1$), and a 0.14% probability of a negative difference". Such a statement has no frequentist counterpart.

The question of replication of observations. The Bayesian inference offers a direct and very intuitive solution. Given the performed experiment, the predictive distribution expresses our state of knowledge about future data. For instance, for an additional experimental unit, "there is a 87.4% probability of a positive difference and a 78.8% probability of a difference exceeding half one hour", and for a future sample of size 10, "there is a 99.1% probability of a positive difference and a 95.9% probability of a difference exceeding half an hour".

Power and sample size: Bayesian data planning and monitoring. "An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results." (Berry, 1991, p.81). Bayesian predictive procedures give users a very appealing method to answer essential questions such as: "how big should be the experiment to have a reasonable chance of demonstrating a given conclusion?"; "given the current data, what is the chance that the final result will be in some sense conclusive, or on the

contrary inconclusive?" These questions are unconditional in that they require consideration of all possible value of parameters. Whereas traditional frequentist practice does not address these questions, predictive probabilities give them direct and natural answer.

In particular, from a pilot study, the predictive probabilities on credibility limits give a useful summary to help in the choice of the sample size of an experiment. If the data from the pilot study are included in the final analysis, final results for the whole data can be predicted as well (Lecoutre, 2001). Predictive procedures can also be used to aid the decision to abandon an experiment if the predictive probability appears poor. Some relevant references are Berry (1991), Lecoutre, Derzko and Grouin (1995), Joseph and Bélisle (1997), Dignam *et al.*, (1998), Johns and Andersen (1999), Lecoutre (2001), Lecoutre, Mabika and Derzko (2002).

Introducing "informative" priors. If the use of noninformative priors has a privileged status in order to gain "public use" statements, other Bayesian techniques also have an important role to play in experimental investigations. They are ideally suited for combining information from several studies and therefore planning a series of experiments. Realistic uses of these techniques have been proposed. When a fiducial Bayesian analysis suggests a given conclusion, various prior distributions expressing results from other experiments or subjective opinions from specific, well-informed individuals ("experts"), which whether *skeptical* or *enthusiastic*, can be investigated to assess the robustness of conclusions (see in particular Spiegelhalter, Freedman and Parmar, 1994). With regard to scientists' need for objectivity, it could be argued with Dickey (1986, p.135) that "an objective scientific report is a report of the whole prior-to-posterior mapping of a relevant range of prior probability distributions, keyed to meaningful uncertainty interpretations".

3.4 The Feasibility of Bayesian Methods

We especially developed Bayesian methods in the analysis of variance framework, which is an issue of particular importance for experimental data analysis. Experimental investigations frequently involve complex designs, especially repeated-measures designs. Bayesian procedures have been developed on the subject, but they are generally thought difficult to implement and not included in the commonly available computer packages. As a consequence the possibility of teaching them is still largely questionable for many statistical teachers.

A simple way to deal with the complexity of experimental designs it is to use the *specific analysis approach*. Roughly speaking, a specific analysis for a particular effect consists in handling only data that are *relevant for it*. Most of-

ten, the design structure of these relevant data is much simpler that the original design structure, and the number of "nuisance" parameters involved in the specific inference is drastically reduced. Consequently, in the Bayesian framework, relatively *elementary* procedures can be applied and *realistic prior* distributions can be investigated. Furthermore, necessary and minimal assumptions specific to each particular inference are made explicit. When these assumptions are under suspicion, alternative procedures can be easily envisaged: for instance we can do a transformation of the relevant data, or again use solutions that do not assume the equality of variances, etc. Thus, the advantages of the specific analysis approach over the conventional general model approach appear overwhelming both for the feasibility and the understanding of procedures.

Further justifications can be found in Rouanet and Lecoutre (1983) (see also Lecoutre, 1984 and Rouanet, 1996). Note that the interest of the specific analysis approach to analysis of variance is often implicitly recognized. In this way, Hand and Taylor (1987) suggested systematically deriving relevant data before using commonly available computer packages. In a more particular context Jones and Kenward (1989) developed a "simple and robust analysis for two-group dual designs" (page 160) which is typically a specific analysis.

Three decisive advantages of the specific analysis approach can be stressed. (1) All the traditional analysis of variance procedures can be derived as a direct extension of the basic procedures used in descriptive statistics (means, standard deviations) and inferential statistics (Student's t tests). (2) Complex designs involving several factors can easily be handled; in particular, the exact validity assumptions for each inference can be made explicit and comprehensible. (3) Bayesian procedures become straightforward to implement.

Statistical computer programs based on the specific inference approach have been developed (Lecoutre and Poitevineau, 1992; Lecoutre, 1996; Lecoutre and Poitevineau, 2005[2]). They incorporate both traditional frequentist practices (significance tests, confidence intervals) and Bayesian procedures (non informative and conjugate priors). These procedures are applicable to general experimental designs (in particular, repeated measures designs), balanced or not balanced, with univariate or multivariate data, and covariables.

Other packages designed to teach or learn elementary Bayesian Statistical inference are *First Bayes* (O'Hagan, 1996)[3] and a package of *Minitab* macros (Albert, 1996).

I have restricted here my presentation to the analysis of variance framework; however similar materials are also available for inferences about propor-

[2]Retrived March 9, 2006 from http://www-rocq.inria.fr/axis/modulad/logiciels.htm#lepac.

[3]O'Hagan, A. (1996). *First Bayes* [Teaching package for elementary Bayesian Statistics]. Retrieved march 9, 2006 from http://www.shef.ac.uk/∼st1ao/1b.html.

tions (Lecoutre, Derzko and Grouin, 1995; Bernard, 2000; Lecoutre and Charron, 2000).

4. Training Students and Researchers in Bayesian Methods

"*It is their straightforward, natural approach to inference that makes them* [Bayesian methods] *so attractive*" (Schmitt, 1969, preface)

In 1976 Jaynes wrote "As a teacher, I therefore feel that to continue the time honoured practice – still in effect in many schools – of teaching pure orthodox statistics to students, with only a passing sneer at Bayes and Laplace, is to perpetuate a tragic error which has already wasted thousands of man-years of our fines mathematical talent in pursuit of false goals. If this talent had been directed toward understanding Laplace's contributions and learning how to use them properly, statistical practice would be far more advanced than it is." (Jaynes, 1976, p.256). It would be folly to perpetuate this error! For more than twenty-five years now, with my colleagues we have gradually introduce Bayesian methods in courses and seminars for audiences of various backgrounds, especially in psychology. Our statistical teaching and consulting experience revealed us that these methods were far more intuitive and much closer to the thinking of scientists than frequentist procedures. So we completely disagree with Moore (1997) who claimed that "Bayesian reasoning is considerably more difficult to assimilate than the reasoning of standard inference".

4.1 Teaching strategy

Since experimental publications are full of significance tests, students and researchers are (and will be again in the future) constantly confronted to their use. NHST is so an integral part of scientists' behavior and of experimental teaching that its misuses and abuses should not be discontinued by flinging it out of the window, even if I completely agree with Rozeboom (1997, p.335) that NHST is "surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students". This reality cannot be ignored, and it is a challenge for the teachers of statistics to introduce Bayesian inference without discarding, neither NHST nor the "official" guidelines that tend to supplant it by confidence intervals. So I argue that the sole effective strategy is *a smooth transition towards the Bayesian paradigm* (see Lecoutre, Lecoutre and Poitevineau, 2001).

The suggested teaching strategy is to introduce Bayesian methods as follows. (1) To present natural *fiducial Bayesian interpretations* of NHST outcomes to call attention about their shortcomings. (2) To create as a result of this the need for *a change of emphasis in the presentation and interpretation* of results.

(3) Finally to equip students with a real possibility of *thinking sensibly about statistical inference* problems and behaving in a more reasonable manner.

From an interactive use of our computer programs, a very limited set of preliminary notions is needed to introduce basic ANOVA procedures, that is inferences about one degree of freedom effects in complex designs. The possibility of applying Bayesian methods in the context of realistic complex experimental designs is an essential requirement for motivating students and researchers. The attention can be concentrated about the basic principles and the practical meaning of procedures. As a consequence, the principles of advanced techniques can be more easily understood, independently of their mathematical difficulty.

4.2 First example: student data

It is remarkable to notice that the Student's example presented in Section 3.3 was a typical application of the specific analysis approach. The basic data were for each of the n=10 patients the difference between the two "additional hour's sleep gained by the use of hyoscyamine hydobromide [an hypnotic]", the hour's sleep being measured without drug and after treatment with either (1) "dextro hyoscyamine hydobromide" or (2) "laevo hyoscyamine hydobromide" (note that they already were derived data). The Student's analysis is a typical example of specific inference: it only involves the elementary inference about a normal mean.

In the same way, we can apply to the data in Table 1 the elementary Bayesian inference about a normal mean, with only two parameters, the population mean difference δ and the standard deviation σ. Assuming the usual noninformative prior, the posterior (fiducial Bayesian) distribution of δ is a generalized (or scaled) t distribution. It is centered on the mean observed difference $d = +1.58$ and has a scale factor $e = s/\sqrt{n} = 0.39$. The distribution has the same degrees of freedom q=9 as the t test.

This is written $\delta \sim d + et_q$, or again $\delta \sim t_q(d, e^2)$ – hence here $\delta \sim t_9(+1.58, 0.39^2)$ – by analogy with the normal distribution (note that this distribution must not be confused with the *noncentral* t distribution, familiar to power analysts). The scale factor e is the denominator of the usual t test statistic, that is $e = d/t$ (assuming $d \neq 0$). In consequence, the fiducial Bayesian distribution of δ can be directly derived from t=+4.06. This result brings to the fore the fundamental property of the t test statistic of being an estimate of the experimental accuracy, *conditionally on the observed value* d. More precisely, $(d/t)^2$ estimates the sampling error variance of d.

Resorting to computers solves the technical problems involved in the use of Bayesian distributions. This gives the students an attractive and intuitive way of understanding the impact of sample sizes, data and prior distributions. The posterior distribution can be investigated by means of visual display. The fiducial

Bayesian interpretation of usual significance tests is made explicit. The credibility limits for a given probability (or guarantee), or conversely the probability of a given interval can be computed.

An important aspect of statistical inference is making predictions. Again, the Bayesian inference offers a direct and very intuitive solution. For instance, what can be said about the value of the difference d' that would be observed for new data? The predictive distribution for d' in a future sample of size n' is naturally more scattered than the distribution of δ relative to the population (this is all the more true since the size of the new sample is smaller). Thus the fiducial Bayesian (posterior) predictive distribution for d', given the value d observed in the available data, is again a generalized t distribution (naturally centered on d), $d' \sim t_q(d, e^2 + e'^2)$, where $e' = s/\sqrt{n'}$. In fact, the uncertainty about δ given the available data (reflected by e^2) is added to the uncertainty about the results of the future sample when δ is known (reflected by e'^2). Given the Student's data, the predictive distribution is $d' \sim t_9(+1.58, 1.29^2)$ for a future experimental unit ($n' = 1$) and $d' \sim t_9(+1.58, 0.55^2)$ for a replication with the same sample size ($e' = e$).

4.3 Second example: reaction time experiment

As an illustration of a more complex design, let us consider the following example, derived from Holender and Bertelson (1975). In a psychological experiment, the subject must react to a signal. The experimental design involves two crossed repeated factors: Factor A (signal frequency) with two levels ($a1$: frequent and $a2$: rare), and Factor B (foreperiod duration), with two levels ($b1$: short and $b2$: long). The main research hypothesis is a null (or about null) interaction effect between factors A and B (*additive model*). The $n = 12$ subjects are divided into three groups of four subjects each. The data treated here and reported in Table 2 are reaction times in ms (averaged over trials). They have been previously analysed in detail with Bayesian methods in Rouanet and Lecoutre (1983), Rouanet (1996) and Lecoutre and Derzko (2001). I will focus here on the technical aspects of the specific analysis approach for one degree of freedom sources of variations, but this approach can be easily generalized to several *df* sources.

Here the basic data consists of three "groups" and four "occasions" of measure. Since A and B are both two-level factors, their interaction can be represented by a single contrast among the four occasions. Let us consider the contrast with coefficients $[w_o]_{o \in O} = [+1 - 1 - 1 + 1]$. The coefficients $[w_o]$ are called *coefficients of derivation upon occasions*. The derived relevant data for interaction consist of the twelve individual interaction effects reported in Table 2. They constitute a simple (balanced) one-way layout and the interaction effect amounts

to the overall mean δ. This mean is given by the *coefficients of derivation upon groups* $[v_g]_{g\in G} = [1/3\ 1/3\ 1/3]$.

Table 2: Reaction time experiment: basic data and relevant data for interaction and for groups comparisons

						derived individual data	
group	subject	$a1b1$	$a2b1$	$a1b2$	$a2b2$	interaction effect	mean
g1	1	387	435	416	473	+9	427.75
	2	321	336	343	368	+10	342.00
	3	333	362	358	390	+3	360.75
	4	344	430	352	393	-45	379.75
	mean					$d_{g1} = -5.75$	$d_{g1} = 377.56$ ms
						$s_{g1} = 26.35$ ms	$s_{g1} = 36.84$ ms
g2	5	368	432	432	504	+8	434.00
	6	357	367	394	411	+7	382.25
	7	336	346	340	421	+71	360.75
	8	387	454	438	496	-9	443.75
	mean					$d_{g2} = +19.25$ ms	$d_{g2} = 405.19$ ms
						$s_{g2} = 35.37$ ms	$s_{g2} = 40.08$ ms
g3	9	345	408	417	479	-1	412.25
	10	358	389	372	407	+4	381.50
	11	317	375	341	392	-7	356.25
	12	386	510	464	513	-75	468.25
	mean					$d_{g3} = -19.75$ ms	$d_{g3} = 404.56$ ms
						$s_{g3} = 37.11$ ms	$s_{g3} = 48.24$ ms
	mean	353.3	403.7	388.9	437.3	$d = -2.08$ ms	$d = 395.71$ ms
						$s = 33.28$ ms	$s = 41.99$ ms

As a general result, a one *df* effect can be tested from the t statistic $t = d/e = -0.217$, where $e = bs = 9.61$ is precisely the scale factor of the fiducial Bayesian distribution. The constant b depends on the coefficients of derivation upon groups v_g and on the group sizes f_g (here $f_{g1} = f_{g2} = f_{g3} = 4$): $b^2 = \sum(v_g^2/f_g) = 1/12 = 0.289^2$. The *within group* variance $s^2 = 33.28^2$ is the mean of the group variances weighted by their degrees of freedom $f_g - 1$. In the case of unequal group sizes we could consider either the unweighted mean or the weighted mean, given respectively by the coefficients $[v_g]_{g\in G} = [1/3\ 1/3\ 1/3]$ (*unweighted*) and $[v_g]_{g\in G} = [f_{g1}/12\ f_{g2}/12\ f_{g3}/12]$ (*weighted*).

The following general results ensure the link with the traditional ANOVA

procedures. The two mean squares of the usual ANOVA F ratio,

$$F = MS_{A.B}/MS_{S(G).A.B} = 0.047,$$

are respectively proportional to d^2 and s^2: $MS_{A.B} = (d/(ab))^2 = 13.02$ and $MS_{S(G).A.B} = (s/a)^2 = 276.84$. The constant a only depends on the coefficients of derivation upon occasions w_o: $a^2 = \sum w_o^2 = 4$. All these formulae are made explicit in our computer programs. With these notations, all inferential (frequentist and Bayesian) procedures are simply modeled on the inference on a normal mean.

Any one *df* source of variation of interest can be analyzed in the same way. Suppose for instance that group $g3$ is a *control* group; then we may plan to decompose some effects involving factor G according the following two contrasts: $g2, g1$ (opposing $g2$ and $g1$) and $g3, g1\_g2$ (opposing $g3$ on the one hand and $g1$ and $g2$ on the other hand). The specific analysis of these two contrasts involves as relevant data the twelve individual means reported in Table 2. The coefficients of derivation upon occasions are $[w_o] = [1/4\ 1/4\ 1/4\ 1/4]$ ($a^2 = 1/4$) and we consider for the derived data the two (orthogonal) contrasts between groups with the respective coefficients $[-1\ +1\ 0]$ ($b^2 = 1/2$) and $[-1/2\ -1/2\ +1]$ ($b^2 = 1/2$). From the relevant data for interaction, we can again analyze the interactions between $A.B$ and these two contrasts. Table 3 gives a summary of the specific analyses of all sources of variations.

Table 3: Reaction time experiment: Summary table of specific analyses

Between subjects	$[w_o]$	$[v_g]$	a	b	d	s	$e{=}bs$
$g2, g1$	aaaa	egf	0.5	0.7071	+27.63	41.99	29.69
$g3, g1\_g2$	aaaa	ccf	0.5	0.6124	+13.19	41.99	25.71
Within subjects	$[w_o]$	$[v_g]$	a	b	d	s	$e{=}bs$
$a2, a1$	cbcb	ddd	1	0.2887	+49.38	22.26	6.42
$a2, a1.g2, g1$	cbcb	egf	1	0.7071	+5.75	22.26	15.74
$a2, a1.g3, g1\_g2$	cbcb	ccf	1	0.6124	+14.63	22.26	13.63
$b2, b1$	cbcb	ddd	1	0.2887	+34.63	20.80	6.01
$b2, b1.g2, g1$	ccbb	egf	1	0.7071	+30.50	20.80	14.71
$b2, b1.g3, g1_g2$	ccbb	ccf	1	0.6124	+3.75	20.80	12.74
$A.B$	feef	ddd	2	0.2887	−2.08	33.28	9.61
$A.B.g2, g1$	feef	egf	2	0.7071	+25.00	33.28	23.53
$A.B.g3, g1\_g2$	feef	ccf	2	0.6124	−26.50	33.28	20.38

Codings for the weights: a=+1/4, b=1/2, c=−1/2, d=1/3, e=−1, f=1, g=0.

5. A Challenge for Statistical Instructors

Training students and researchers in Bayesian methods should become an attractive challenge for statistical instructors. It is often claimed that Bayesian methods need new probabilistic concepts, in particular the Bayesian definition of probability, conditional probabilities and Bayes' formula. However, since most people use "inverse probability" statements to interpret NHST and confidence intervals, these notions are already – at least implicitly – involved in frequentist methods. Which is simply required for teaching the Bayesian approach is a very natural shift of emphasis about these concepts, showing that they can be used consistently and appropriately in statistical analysis.

5.1 A natural change of emphasis about probabilistic concepts

"[Bayesian analysis provides] *direct probability statements – which are what most people wrongly assume they are getting from conventional statistics*" (Grunkemeier and Payne, 2002, p.1901)

A recent empirical study (Albert, 2003) indicates that students in introductory statistics class are generally confused about the different notions of probabilities. Clearly, teaching NHST and confidence intervals can only add to confusion, since these methods are justified by frequentist arguments and generally (mis)interpreted in Bayesian terms. Ironically these *heretic* interpretations are encouraged by the duplicity of most statistical instructors who tolerate and even use them. For instance Pagano (1990, p.288) describes a 95% confidence interval as "an interval such that the probability is 0.95 that the interval contains the population value". Other authors claim that the "correct" frequentist interpretation they advocate can be expressed as "we can be 95% confident that the population mean is between 114.06 and 119.94" (Kirk, 1982, page 43), "95% confident that θ is below $B(X)$" (Steiger and Fouladi, 1997, p.230) or "we may claim 95% confidence that the population value of multiple R^2 is no lower than .0266" (Smithson, 2001, p.614). It is hard to imagine that students or scientists can understand that "confident" refers here to a frequentist view of probability! So, in a recent paper, Schweder and Hjort (2002) gave the following revealing definitions of probability: "we will distinguish between probability as frequency, termed probability, and *probability as information/uncertainty, termed confidence*" (italics added). After many attempts to teach the "correct" interpretation of frequentist procedures, I completely agree with Freeman (1993) that in these attempts "we are fighting a losing battle".

Regarding conditional probability and Bayes' formula, the traditional teaching of frequentist procedures is also misleading. This is especially revealed by the fact that even experienced researchers frequently confused "the [*conditional*]

probability of making a Type I error if the null hypothesis is true" and "the *marginal* probability of making a Type I error". So Azar (1999)[4] wrote: "[a significant result] indicates that the chances of the finding being random is only 5 percent or less"; this statement was later commented by Bakeman (1999)[5] as "a misunderstanding that generations of instructors of statistics clearly have failed to eradicate". This can be due to the fact that little or no emphasis is placed on conditional probabilities in most of the frequentist presentations. For instance, standard statistical textbooks speak about "the probability of making a Type I [Type 2] error" by omitting the conditional argument "given H_0 [H_1]" (see e.g., Kirk, 1982, pages 36-37). I believe with Berry (1997) that conditional probabilities are intuitive for many people. Also, Bayes' formula is easily understood if it is introduced from contingency tables with probabilities interpreted as frequencies so that prior probabilities can be supposed exactly known (see Box and Tiao, 1973, p.12).

Considerable difficulties are due to the mysterious and unrealistic use of the sampling distribution for justifying NHST and confidence intervals. Frequent questions asked by students show us that this use is counterintuitive: "why must one calculate the probability of samples that have not been observed?"; "why one considers the probability of samples outcomes that are more extreme than the one observed?"; etc. Such difficulties are not encountered with the Bayesian inference: the posterior distribution, being conditional on data, only involves the sampling probability of the data *in hand*, via the likelihood function that writes the sampling distribution in the "natural order".

5.2 The Bayesian approach gives tools to overcome usual difficulties

"*I stopped teaching frequentist methods when I decided that they could not be learned*" (Berry, 1997).

There are hardly – if not Bayesian – *intuitive* justifications of frequentist procedures. On the contrary, with the Bayesian approach, intuitive justifications and interpretations of procedures can be given, so that the level of mathematical justifications can be easily adapted to the students state of knowledge. So it can be argued with Albert (1995[6], 1997) and Berry (1997) that elementary Bayesian inference can be taught effectively to undergraduate students and that students benefit greatly from such instruction. Moreover, an empirical understanding of probability concepts is gained by applying Bayesian procedures, especially with the help of computer programs.

Our experience with Bayesian methods is that they allow students to over-

[4]Retrieved March 9, 2006 from http://www.apa.org/monitor/may99/task.html.
[5]Retrieved March 9, 2006 from http://www.apa.org/monitor/julaug99/letters.html.
[6]Retrieved March 9, 2006 from http://www.amstat.org/publications/jse/v3n3/albert.html.

come usual difficulties encountered with the frequentist approach. Of course, the following list is not exhaustive and empirical studies for asserting our conclusions should be welcome.

It can be hard for students to distinguish a parameter, such as a population mean, from the observed mean statistic computed from a sample. The two notions of posterior distribution and predictive distribution of future data, given available data, are useful tools to give students an understanding of this essential distinction. Moreover, the predictive distribution can be used to give, as limiting cases: (1) the sampling distribution of a statistic when the prior distribution tends to a point distribution ("known parameter"); (2) the posterior distribution when the sample size of the future data tends to infinity (the parameter can be seen as the observed statistic in a future sample of very large size).

Moreover, the notions of posterior and predictive distributions, being fundamental tools for a better understanding of sample fluctuations, allow the students to be aware of misconceptions about the replication of experiments. Indeed, many people overestimate the probability of repeating a significant result (Tversky and Kahneman, 1971; Lecoutre and Rouanet, 1993). Similar misconceptions are encountered with confidence intervals. An empirical study (Cumming *et al.*, 2004) suggested that many "leading researchers" in psychology, behavioural neuroscience, and medicine "hold the confidence level misconception that a 95% CI will on average capture 95% of replication means", underestimating the extent that future replications will vary.

An important difficulty with the logic of NHST is that it requires that the hypothesis to be demonstrated should be the alternative hypothesis. Of course this artifice can be completely avoided with the Bayesian approach, which provide direct answers to the right questions: "what is the probability that the difference between two means is large?"; "what is the probability that the difference (in absolute value) is small?"; "given the current inconclusive partial data, what is the chance that the final result will be conclusive?"; etc.

Using the fiducial Bayesian interpretations of significance tests and confidence intervals in the natural language of probabilities about unknown effects comes quite naturally to students. In return the common misuses and abuses of NHST appear to be more clearly understood. In particular students become quickly alerted that nonsignificant results cannot be interpreted as "proof of no effect". I completely agree with Berry (1997) who ironically concludes that students exposed only to a Bayesian approach "come to understand the frequentist concepts of confidence intervals and P values better than do students exposed only to a frequentist approach".

An important objective of statistical teaching is to prepare students to read experimental publications. For the reasons exposed above, with the Bayesian

approach students are well equipped for an intelligent and critical reading. In fact, the Bayesian approach fits in better than the frequentist approach with the usual way of reporting experimental results, which seldom involves explicitly the basic concepts of the NHST reasoning (null hypothesis, α level ...).

By interactively investigating various prior distributions and contrasting the resulting posterior with the fiducial Bayesian solution, students can gain understanding and intuition about the relative roles of sample size, data and external information. Investigating predictive distributions by varying the respective sample sizes of the available and future data is also useful to give students an intuitive understanding of the role of sample size.

5.3 Some possible difficulties with the Bayesian approach

The most often denounced difficulties with the Bayesian approach lie in the elicitation of the prior distribution. Berry (1997) places emphasis on the fact that prior and posterior Bayesian distributions are subjective and forces students to assess their prior probabilities, while recognizing the difficulties of this task ("they don't like it"). At least the role of subjective probability should be clarified (D'Agostini, 1999).

However, *insofar as experimental data analysis is concerned*, I do not think that it is a good strategy to draw the attention of students (or researchers) on an approach that does not answer their expectations (see Section 2.1). So, we always avoid – at least *in a first stage* – the issue of assessing a "subjective" prior distribution and focus our teaching on the fiducial Bayesian procedures. Once that students will become familiarized with their use and interpretation, there are appealing ways to introduce "informative" prior distributions at a later stage. In particular, students generally find attractive to investigate the impact of a *handicap* ("skeptical") prior and to examine if the data give sufficient evidence to counterbalance it. Priors that express the results of previous experiments are also generally well-accepted. Finally, on can show that the elicitation of prior opinions from "experts" in the field can be useful in some studies, but it must be emphasized that this needs appropriate techniques (see for an example in clinical trials Tan *et al.*, 2003).

Other difficulties can be due to confusions with the frequentist interpretations. For instance, some students erroneously conclude from the posterior distribution that the *observed* difference – not the population difference – is large, which can be due to a confusion with the NHST reasoning (a result is significant if the observed difference is "in a sense" large). A possibility is not to teach frequentist methods (Berry, 1997). However, in the current context, this is hardly a realistic attitude. An alternative line of attack is to use the combinatorial (or set-theoretic) inference approach suggested by Rouanet and Bert (2000) (see also Rouanet, Bernard

and Lecoutre, 1986; Rouanet, Bernard and Le Roux, 1990). Roughly speaking, this approach consists of ruling out the "randomness" character of the concept of sample and to replace probabilistic formulations by formulations in terms of "proportions of samples". The teaching motivation is to allow students to learn the *computational aspects* of frequentist inference procedures without being prematurely concerned with the conceptual difficulties of probabilistic concepts. Thus, the probabilistic formulations – and in particular the *interpretation* of frequentist procedures – are reserved to the Bayesian approach, minimizing possible source of confusions.

6. Conclusion

"*It could be argued that since most physicians use statement A* [the probability the true mean value is in the interval is 95%] *to describe 'confidence' intervals, what they really want are 'probability' intervals. Since to get them they must use Bayesian methods, then they are really Bayesians at heart!*" (Grunkemeier and Payne, 2002, p.1904)

Nowadays Bayesian routine methods for the familiar situations of experimental data analysis are easy to implement. They fulfill the requirements of scientists and they fit in better with their spontaneous interpretations of data than frequentist procedures. So they can be taught to non-statistician students and researchers in intuitively appealing form. Using the fiducial Bayesian (using noninformative priors) interpretations of significance tests and confidence intervals in the natural language of probabilities about unknown effects comes quite spontaneously to students. In return the Bayesian approach bypasses usual difficulties encountered with frequentist procedures, and in particular the common misuses and abuses of NHST are more clearly understood. Users' attention can be focused to more appropriate strategies such as consideration of the practical significance of results and the replication of experiments.

References

Albert, J. (1995). Teaching inference about proportions using Bayes and discrete models. *Journal of Statistics Education* **3**.

Albert, J. (1996). *Bayesian Computation Using Minitab*. Wadsworth Publishing Company.

Albert, J. (1997). Teaching Bayes' rule: a data-oriented approach. *The American Statistician* **51**, 247-253.

Albert, J. (2003). College students' conceptions of probability. *The American Statistician* **57**, 37-45.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th edition). Author, Washington, DC.

Amorim, M. A., Glasauer, S., Corpinot, K. and Berthoz, A. (1997). Updating an object's orientation and location during nonvisual navigation: A comparison between two processing modes. *Perception and Psychophysics* **59**, 404-418.

Amorim, M. -A., Loomis, J. M. and Fukusima, S. S. (1998). Reproduction of object shape is more accurate without the continued availability of visual information. *Perception* **27**, 69-86.

Amorim, M.-A. and Stucchi, N. (1997). Viewer- and object-centered mental explorations of an imagined environment are not equivalent. *Cognitive Brain Research* **5**, 229-239.

Amorim, M. -A., Trumbore, B. and Chogyen, P. L. (2000). Cognitive repositioning inside a desktop VE: The constraints introduced by first- versus third-person imagery and mental representation richness. *Presence: Teleoperators and Virtual Environments* **9**, 165-186.

Azar, B. (1999). APA statistics task force prepares to release recommendations for public comment. *APA Monitor Online* **30**.

Bakeman, R. (1999). Statistical matters (letter). *APA Monitor Online* **30**(7).

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* **33**, 526-542.

Bernard, J.-M. (2000). Bayesian inference for categorized data. In *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (Edited by H. Rouanet, J. -M., Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre and B. Le Roux, (2nd edition)), 159-226. Peter Lang.

Berry, D. A. (1991). Experimental design for drug development: A Bayesian approach. *Journal of Biopharmaceutical Statististics* **1**, 81-101.

Berry, D. A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician* **51**, 241-246.

Boring, E. G. (1919). Mathematical versus scientific significance. *Psychological Bulletin* **16**, 335-338.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison Wesley.

Ciancia, F., Maitte, M., Honoré, J., Lecoutre, B. and Coquery, J. -M. (1988). Orientation of attention and sensory gating: An evoked potential and RT study in cat. *Experimental Neurology* **100**, 274-287.

Clément, E. and Richard, J.-F. (1997). Knowledge of domain effects in problem representation: the case of Tower of Hanoi isomorphs. *Thinking and Reasoning* **3**, 133-157.

Cumming, G., Williams, J. and Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics* **3**, 299-311.

D'Agostini, G. (1999). Teaching statistics in the physics curriculum. Unifying and clarifying role of subjective probability. *American Journal of Physics*, **67**, 1260-1268.

Desperati, C. and Stucchi, N. (1995). The role of eye-movements. *Experimental Brain Research* **105**, 254-260.

Dickey J. M. (1986). Discussion of Racine, A., Grieve, A. P., Flühler, H. and Smith, A. F. M., Bayesian methods in practice: Experiences in the pharmaceutical industry. *Applied Statistics* **35**, 93-150.

Dignam, J., Bryant, J., Wieand, H. S., Fisher, B. and Wolmark, N. (1998). Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Controlled Clinical Trials* **19**, 575-588.

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

Fisher, R. A. (1990/1925). *Statistical Methods for Research Workers*. Oliver and Boyd, London. (Reprint, 14th edition, *in* Fisher, 1990).

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.

Gordon, H. R. D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research* **26**, 244-271.

Grunkemeier, G. L. and Payne, N. (2002). Bayesian analysis: A new statistical paradigm for new technology. *The Annals of Thoracic Surgery* **74**, 1901-1908.

Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research* **7**.

Hand, D. J. and Taylor, C. (1987). *Multivariate Analysis of Variances and Repeated Measures: A practical Approach for Behavioural Scientists*. Chapman and Hall.

Hoc, J. -M. (1996). Operator expertise and verbal reports on temporal data. *Ergonomics* **39**, 811-825.

Hoc, J. -M. and Leplat, J. (1983). Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies* **18**, 283-306.

Holender, D. and Bertelson, P. (1975). Selective preparation and time uncertainty. *Acta Psychologica* **39**, 193-203.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. 2* (Rdited by W. L. Harper and C. A. Hooker), 175-257, D. Reidel, Dordrecht.

Jeffreys, H. (1998). *Theory of Probability* (3rd edition). Clarendon.

Johns, D. and Andersen, J. S. (1999). Use of predictive probabilities in phase II and phase III clinical trials. *Journal of Biopharmaceutical Statistics* **9**, 67-79.

Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Cross-over Trials*. Chapman and Hall.

Joseph, L. and Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46**, 209-226.

Kirk, R. E. (1982). *Experimental Design. Procedures for the Behavioral Sciences*. Brooks /Cole.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement* **61**, 213-218.

Laplace, P.-S. (1986/1825). *Essai Philosophique sur les Probabilités*. Christian Bourgois (English translation: *A Philosophical Essay on Probability*, 1952, Dover, New York).

Lecoutre, B. (1984). *L'Analyse Bayésienne des Comparaisons.* Presses Universitaires de Lille.

Lecoutre, B. (1996). *Traitement statistique des données expérimentales: des pratiques traditionnelles aux pratiques bayésiennes*. DECISIA Editions.

Lecoutre, B. (2000). From significance tests to fiducial Bayesian inference. In *New ways in statistical methodology From significance tests to Bayesian inference* (Edited by H. Rouanet, J. -M. Bernard, M. -C. Bert, B. Lecoutre, M. -P. Lecoutre and B. Le Roux, 2nd edition), 123-157, Peter Lang.

Lecoutre B. (2001). Bayesian predictive procedure for designing and monitoring experiments. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, 301-310, Office for Official Publications of the European Communities, Luxembourg.

Lecoutre, B. and Charron, C. (2000). Bayesian procedures for prediction analysis of implication hypotheses in 2×2 contingency tables. *Journal of Educational and Behavioral Statistics* **25**, 185-201.

Lecoutre, B. and Derzko, G. (2001). Asserting the smallness of effects in ANOVA. *Methods of Psychological Research* **6**, 1-32.

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

Lecoutre, B., Lecoutre, M. -P. and Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review* **69**, 399-418.

Lecoutre, B., Mabika, B. and Derzko, G. (2002). Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups: a Bayesian approach with Weibull modeling illustrated. *Statistics in Medicine* **21**, 663-674.

Lecoutre, B. and Poitevineau, J. (1992). PAC (*Programme d'Analyse des Comparaisons*): *Guide d'utilisation et manuel de référence*. CISIA-CERESTA.

Lecoutre, B. and Poitevineau, J. (2000). Aller au delà des tests de signification traditionnels: vers de nouvelles normes de publication. *L'Année Psychologique* **100**, 683-713.

Lecoutre, M. -P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics* **23**, 557-568.

Lecoutre, M. -P. (2000). And ... What about the researcher's point of view. In *New ways in statistical methodology: From significance tests to Bayesian inference* (Edited by H. Rouanet, J. -M. Bernard, M. -C. Bert, B. Lecoutre, M. -P. Lecoutre and B. Le Roux, 2nd edition), 65-95. Peter Lang.

Lecoutre, M.-P., Poitevineau, J. and Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology* **38**, 37-45.

Lecoutre, M.-P., Clément, E. and Lecoutre, B. (2004). Failure to construct and transfer correct representations across probability problems. *Psychological Reports* **94**, 151-162.

Lecoutre, M.-P. and Poitevineau, B. (2005). Le Logiciel "LePAC". *La Revue de Modulad* **33**.

Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology* **38**, 37-45.

Lecoutre, M.-P. and Rouanet H. (1993). Predictive judgments in situations of statistical analysis. *Organizational Behavior and Human Decision Processes* **54**, 45-56.

Lee, P. (1997). *Bayesian Statistics: An Introduction* (2nd edition). Oxford University Press.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics* **15**, 22-25.

Mittag, K. C. and Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher* **29**, 14-20.

Moore, D.S. (1997). Bayes for Beginners? Some Pedagogical Questions. In S. Panchapakesan and N. Balakrishnan (eds.), *Advances in Statistical Decision Theory*, 3-17, Birkhäuser.

Morrison, D. E. and Henkel, R. E. (Eds.) (1970). *The Significance Test Controversy – A Reader*. Butterwoths.

Nelson, N., Rosenthal, R. and Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist* **41**, 1299-1301.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics* **45**, 401-410.

Oakes, M. (1986). *Statistical Inference: A Commentary for The Social and Behavioural Sciences*. Wiley.

Pagano, R. R. (1990). *Understanding Statistics in the Behavioral Sciences* (3rd edition). West.

Phillips, L. D. (1973). *Bayesian Statistics for Social Scientists*. Nelson.

Poitevineau J. (1998). *Mehodologie de l'analyse des données exprimentales – Etude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive*. Thése de doctorat de psychologie, Université de Rouen (France).

Poitevineau, J. and Lecoutre, B. (2001). The interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated. *Psychonomic Bulletin and Review* **8**, 847-850.

Robinson, D. H. and Wainer, H. (2002). On the past and future of Null Hypothesis Significance Testing. *Journal of Wildlife Management* **66**, 263-271.

Rosenthal, R. and Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology* **55**, 33-38.

Rosnow, R.L. and Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods* **1**, 331-340.

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

Rouanet, H. (2000). Statistical practice revisited. In *New ways in statistical methodology: From significance tests to Bayesian inference* (Edited by H. Rouanet, J. -M. Bernard, M. -C. Bert, B. Lecoutre, M. -P. Lecoutre and B. Le Roux, 2nd edition), 29-64. Peter Lang.

Rouanet, H., Bernard, J. -M. and Lecoutre, B. (1986). Non-probabilistic statistical inference: A set theoretic approach. *The American Statistician* **40**, 60-65.

Rouanet, H., Bernard, J.-M. and Leroux, B. (1990). *Statistique en Sciences Humaines: Analyse Inductive des Données*. Dunod.

Rouanet, H. and Bert, M.-C. (2000). Introduction to combinatorial inference. In *New ways in statistical methodology: From significance tests to Bayesian inference* (Edited by H. Rouanet, J. -M. Bernard, M. -C. Bert, B. Lecoutre, M. -P. Lecoutre and B. Le Roux, 2nd edition), 97-122. Peter Lang.

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.

Rozeboom, W. W (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin* **57**, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In *What If There Were No Significance Tests?* (EDited by L. L. Harlow, S. A. Mulaik and J. H. Steiger), 335-392, Erlbaum.

Savage, L. (1954). *The Foundations of Statistical Inference*. John Wiley and Sons.

Schmitt, S. A. (1969). *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Addison Wesley.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309-332.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement* **61**, 605-632.

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

Steiger, J. H. and Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In *What If There Were No Significance Tests?* (Edited by L. L. Harlow, S. A. Mulaik and J. H. Steiger), 221-257. Erlbaum.

Student (1908). The probable error of a mean. *Biometrika* **6**, 1-25.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* **76**, 237-251.

Tyler, R. (1931). What is statistical significance? *Educational Research Bulletin* **10**, 118-142.

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.

Zuckerman, M., Hodgins, H., Zuckerman, A. and Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science* **4**, 49-53

Received September 3, 2004; accepted November 25, 2004.

Bruno Lecoutre
ERIS, Laboratoire de Mathématiques Raphaël Salem
UMR 6085 C.N.R.S. et Université de Rouen
Avenue de l'Université, BP 12
76801 Saint-Etienne-du-Rouvray (France)
bruno.lecoutre@univ-rouen.fr