Handbook of Statistics, Vol. 27
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(07)27027-0

The Bayesian Approach to Experimental Data Analysis

Bruno Lecoutre

Abstract

This chapter introduces the conceptual basis of the objective Bayesian approach to experimental data analysis and reviews some of its methodological improvements. The presentation is essentially non-technical and, within this perspective, restricted to relatively simple situations of inference about proportions. Bayesian computations and softwares are also briefly reviewed and some further topics are introduced.

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

Preamble: and if you were a Bayesian without knowing it?

In a popular statistical textbook that claims the goal of "understanding statistics," 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.

If you agree with this statement, or if you feel that it is not the correct interpretation but that it is desirable, you should ask yourselves: "and if I was a Bayesian without knowing it?"

The *correct* 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. Unfortunately, treating the data as random *even after observation* is so strange that this "correct" interpretation does not make sense for most users. Actually, virtually all users interpret frequentist confidence intervals in terms of "a *fixed* interval having a 95% chance of including the true value of interest."

In the same way, many statistical users misinterpret the *p*-values of null hypothesis significance tests as "inverse" probabilities: $1-p$ is "the probability that the alternative hypothesis is true." Even experienced users and experts in statistics (Neyman himself) are not immune from *conceptual* confusions.

> In these conditions [a *p*-value of 1/15], the odds of 14 to 1 that this loss was caused by seeding [of clouds] do not appear negligible to us. (Battan et al., 1969)

After many attempts to rectify these (Bayesian) interpretations of frequentist procedures, I completely agree with Freeman (1993, p. 1446) that in these attempts "we are fighting a losing battle."

> It would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms. (Rouanet, 2000b, p. 54)

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

1. Introduction

Efron (1998, p. 106) wrote

> 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.

I suggest that such a theory is by no means a speculative viewpoint but, on the contrary, is perfectly feasible (see especially, Berger, 2004). It is better suited to the needs of users than frequentist approach and provides scientists with relevant answers to essential questions raised by experimental data analysis.

1.1. What is Bayesian inference for experimental data analysis?

One of the most important objective of controlled clinical trials is to impact on public health, so that their results need to be accepted by a large community of scientists and physicians. For this purpose, null hypothesis significance testing (NHST) has been long conventionally required in most scientific publications for analyzing experimental data. This publication practice dichotomizes each experimental result (significant vs. non-significant) according to the NHST outcome.

But scientists cannot in this way find all the answers to the precise questions posed in experimental investigations, especially in terms of effect size evaluation.

> 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)

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. Within this context, many Bayesians place emphasis on a *subjective* perspective. This can be the reasons why until now scientists have been reluctant to use Bayesian inferential procedures in practice for analyzing their data. It is not surprising that the most common (and easy) criticism of the Bayesian approach by frequentists is the need for prior probabilities. Without dismissing the merits of the decision-theoretic viewpoint, it must be recognized that there is another approach that is just as Bayesian, which was developed by Jeffreys in 1930s (Jeffreys, 1961/1939). 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 *non-informative* prior probabilities are vague distributions that, 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 as a standard.

> A common misconception is that Bayesian analysis is a subjective theory; this is neither true historically nor in practice. The first Bayesians, Bayes (see Bayes (1763)) and Laplace (see Laplace (1812)) performed Bayesian analysis using a constant prior distribution for unknown parameters … (Berger, 2004, p. 3)

1.2. Routine Bayesian methods for experimental data analysis

For more than 25 years now, with other colleagues in France we have worked in order to develop routine Bayesian methods for the most familiar situations encountered in experimental data analysis. These methods can be learned and used as easily, if not more, as the t, F or χ^2 tests. We argued that they offer promising new ways in statistical methodology (Rouanet et al., 2000).

We have especially developed methods based on non-informative priors. In order to promote them, it seemed important to us to give them a more explicit name than "standard," "non-informative" or "reference." Recently, Berger (2004) proposed the name *objective Bayesian analysis*.

> The statistics profession, in general, hurts itself by not using attractive names for its methodologies, and we should start systematically accepting the 'objective Bayes' name before it is co-opted by others. (Berger, 2004, p. 3)

With the same incentive, we argued for the name *fiducial Bayesian* (Lecoutre, 2000; Lecoutre et al., 2001). 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.

An objective (or fiducial) Bayesian analysis has a privileged status in order to gain public use statements. However, this does not preclude using other Bayesian techniques when appropriate.

1.3. The aim of this chapter

The aim of this chapter is to introduce the conceptual basis of objective Bayesian analysis and to illustrate some of its methodological improvements. The presentation will be essentially non-technical and, within this perspective, restricted to simple situations of inference about proportions. A similar presentation for inferences about means in the analysis of variance framework is available elsewhere (Lecoutre, 2006a).

The chapter is divided into four sections. (1) I briefly discuss the frequentist and Bayesian approaches to statistical inference and show the difficulties of the frequentist conception. I conclude that the Bayesian approach is highly desirable, if not unavoidable. (2) Its feasibility is illustrated in detail from a simple illustrative example of inference about a proportion in a clinical trial; basic Bayesian procedures are contrasted with usual frequentist techniques and their advantages are outlined. (3) Other examples of inferences about proportions serve me to show that these basic Bayesian procedures can be straightforward extended to deal with more complex situations. (4) The concluding remarks summarize the main advantages of the Bayesian methodology for experimental data analysis. Bayesian computations and softwares are also briefly reviewed. At last, some further topics are introduced.

The reader interested in more advanced aspects of Bayesian inference, with an emphasis on modeling and computation, is especially referred to the Volume 25 of this series (Dey and Rao, 2005).

2. Frequentist and Bayesian inference

2.1. Two conceptions of probabilities

Nowadays, probability has at least two main definitions (Jaynes, 2003). (1) Probability is the long-run frequency of occurrence of an event, either in a sequence of repeated trials or in an ensemble of "identically" prepared systems. This is the "frequentist" conception of probability, which seems to make probability an observable ("objective") property, existing in the nature independently of us, that should be based on empirical frequencies. (2) Probability is a measure of the degree of belief (or confidence) in the occurrence of an event or in a proposition. This is the "Bayesian" conception of probability.

This dualistic conception was already present in Bernoulli (1713), who clearly recognized the distinction between probability ("degree of certainty") and frequency, deriving the relationship between probability of occurrence in a single trial and frequency of occurrence in a large number of independent trials.

Assigning a frequentist probability to a single-case event is often not obvious, since it requires imagining a reference set of events or a series of repeated experiments in order to get empirical frequencies. Unfortunately, such sets are seldom available for assignment of probabilities in real problems. By contrast, the Bayesian definition is more general: it is not conceptually problematic to assign a probability to a unique event (Savage, 1954; de Finetti, 1974).

> It is beyond any reasonable doubt that for most people, probabilities about single events do make sense even though this sense may be naive and fall short from numerical accuracy. (Rouanet, 2000a, p. 26)

The Bayesian definition fits the meaning of the term probability in everyday language, and so the Bayesian probability theory appears to be much more closely related to how people intuitively reason in the presence of uncertainty.

2.2. Two approaches to statistical inference

The frequentist approach to statistical inference is self-proclaimed *objective* contrary to the Bayesian conception that should be necessary *subjective*. However, the Bayesian definition can clearly serve to describe "objective knowledge," in particular based on symmetry arguments or on frequency data. So Bayesian statistical inference is no less objective than frequentist inference. It is even the contrary in many contexts.

Statistical inference is typically concerned with both known quantities – the observed data – and unknown quantities – the parameters and the data that have not been observed. In the frequentist inference, all probabilities are conditional on parameters that are assumed known. This leads in particular to

- significance tests, where the parameter value of at least one parameter is fixed by hypothesis;
- confidence intervals.

In the Bayesian inference, parameters can also be probabilized. This results in distributions of probabilities that express our uncertainty:

- before observations (they do not depend on data): *prior* probabilities;
- after observations (conditional on data): *posterior* (or *revised*) probabilities;
- about future data: *predictive* probabilities.

As a simple illustration let us consider a finite population of size 20 with a dichotomous variable success/failure and a proportion φ (the *unknown parameter*) of success. A sample of size 5 has been observed, hence these *known data:*

$$0 \quad 0 \quad 0 \quad 1 \quad 0 \qquad f = 1/5$$

The inductive reasoning is fundamentally a generalization from a known quantity (here the data $f = 1/5$) to an unknown quantity (here the parameter φ).

2.3. The frequentist approach: from unknown to known

In the frequentist framework, we have no probabilities and consequently no possible inference. The situation must be reversed, but we have no more probabilities … unless we fix a parameter value. Let us assume, for instance, $\varphi = 0.75$.

Then we get sampling probabilities $\Pr(f|\varphi = 0.75)$ – that is frequencies – involving *imaginary repetitions* of the observations. They can be obtained by simulating repeated drawing of samples of 5 marbles (without replacement) from a box that contains 15 black and 5 white marbles. Alternatively, they can be (exactly) computed from a hypergeometric distribution. These sampling probabilities serve to define a null hypothesis significance test. If the null hypothesis is true ($\varphi = 0.75$), one find in 99.5% of the repetitions a value $f > 1/5$ (the proportion of black marbles in the sample), greater than the observation in hand: the null hypothesis $\varphi = 0.75$ is rejected ("significant test": $p = 0.005$). Note that I do not enter here in the one-sided/two-sided test discussion, which is irrelevant for my purpose.

However, this conclusion is based on the probability of the samples *that have not been observed*, what Jeffreys (1961, Section 7.2) ironically expressed in the following terms:

> If P is small, that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to or greater than the observed set, and the contribution from the actual value is nearly always negligible. What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.

As another example of null hypothesis, let us assume $\varphi = 0.50$. In this case, if the null hypothesis is true ($\varphi = 0.50$), one find in 84.8% of the repetitions a value $f > 1/5$, greater than the observation: the null hypothesis $\varphi = 0.50$ is not rejected by the data in hand. Obviously, *this does not prove that $\varphi = 0.50$*!

Now a frequentist confidence interval can be constructed as the set of possible parameter values that are not rejected by the data. Given the data in hand we get the following 95% confidence interval: [0.05, 0.60]. How to interpret the confidence 95%? The frequentist interpretation is based on the universal statement:

> whatever the fixed value of the parameter is, in 95% (at least) of the repetitions the interval that should be computed includes this value.

But this interpretation is very strange since *it does not involve the data in hand*! It is at least unrealistic, as outlined by Fisher (1990/1973, p. 71):

> Objection has sometimes been made that the method of calculating Confidence Limits by setting an assigned value such as 1% on the frequency of observing 3

or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly the same manner as the value 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation.

2.4. The Bayesian approach: from known to unknown

> As long as we are uncertain about values of parameters, we will fall into the Bayesian camp. (Iversen, 2000)

Let us return to the inductive reasoning, starting from the known data, and adopting a Bayesian viewpoint. We can now use, in addition to sampling probabilities, probabilities that express our uncertainty about all possible values of the parameter. In the Bayesian inference, we consider, not the frequentist probabilities of imaginary samples but the frequentist probabilities of *the observed data* $\Pr(f = 1/5|\varphi)$ for all possible values of the parameter. This is the *likelihood* function that is denoted by

$$\ell(\varphi|\text{data}).$$

We assume prior probabilities $\Pr(\varphi)$ before observations. Then, by a simple product, we get the joint probabilities of the parameter values and the data:

$$\Pr\left(\varphi \text{ and } f = \frac{1}{5}\right) = \Pr\left(f = \frac{1}{5}\middle|\varphi\right) \times \Pr(\varphi) = \ell(\varphi|\text{data}) \times \Pr(\varphi).$$

The sum of the joint probabilities gives the marginal predictive probability of the data, before observation:

$$\Pr\left(f = \frac{1}{5}\right) = \sum_{\varphi} \Pr\left(\varphi \text{ and } f = \frac{1}{5}\right).$$

The result is very intuitive since the predictive probability is a weighted average of the likelihood function, the weights being the prior probabilities.

Finally, we compute the posterior probabilities after observation, by application of the definition of conditional probabilities. The posterior distribution (given by Bayes' theorem) is simply the normalized product of the prior and the likelihood:

$$\Pr\left(\varphi\middle|f = \frac{1}{5}\right) \propto \ell(\varphi|\text{data}) \times \Pr(\varphi) = \frac{\Pr(\varphi \text{ and } f = 1/5)}{\Pr(f = 1/5)}.$$

2.5. The desirability of the Bayesian alternative

We can conclude with Berry (1997):

> Bayesian statistics is difficult in the sense that thinking is difficult.

In fact, it is the frequentist approach that involves considerable difficulties due to the mysterious and unrealistic use of the sampling distribution for justifying null hypothesis significance tests and confidence intervals. As a consequence, even experts in statistics are not immune from *conceptual* confusions about frequentist confidence intervals.

For instance, in a methodological paper, Rosnow and Rosenthal (1996, p. 336) take the example of an observed difference between two means $d = +0.266$. They consider the interval [0, +0.532] whose bounds are the "null hypothesis" (0) and what they call the "counternull value" ($2d = +0.532$), computed as the symmetrical value of 0 with regard to d. They interpret this specific interval [0, +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. 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 $\ell(\varphi|\text{data})$ that writes the sampling distribution in the *natural order:* "from unknown to known."

Moreover, since most people use "inverse probability" statements to interpret NHST and confidence intervals, the Bayesian definition of probability, conditional probabilities and Bayes' formula are already – at least implicitly – involved in the use of frequentist methods. Which is simply required by 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. This makes this approach highly desirable, if not unavoidable.

With the Bayesian inference, intuitive justifications and interpretations of procedures can be given. Moreover, an empirical understanding of probability concepts is gained by applying Bayesian procedures, especially with the help of computer programs.

2.6. Training strategy

The reality of the current use of statistical inference in experimental research cannot be ignored. On the one hand, experimental publications are full of significance tests and students and researchers are (and will be again in the future) constantly confronted to their use. My opinion is that NHST is an inadequate method for experimental data analysis (which has been denounced by the most eminent and most experienced scientists), not because it is an incorrect normative model, just because it does not address the questions that scientific research requires (Lecoutre et al., 2003; Lecoutre, 2006a, 2006b). However, NHST is such an integral part of experimental teaching and scientists' behavior that its misuses and abuses should not be discontinued by flinging it out of the window.

On the one hand, confidence intervals could quickly become a compulsory norm in experimental publications. On the other hand, for many reasons due to their frequentist conception, confidence intervals can hardly be viewed as the ultimate method. In practice, two probabilities can be routinely associated with a specific interval estimate computed from a particular sample.

- The first probability is "the proportion of repeated intervals that contain the parameter." It is usually termed the coverage probability.
- The second probability is the Bayesian "posterior probability that this interval contains the parameter," assuming a non-informative prior distribution.

In the frequentist approach, it is forbidden to use the second probability. On the contrary, in the Bayesian approach, the two probabilities are valid. Moreover, an objective Bayes interval is often "a great frequentist procedure" (Berger, 2004).

As a consequence, it is a challenge for statistical instructors to introduce Bayesian inference without discarding either NHST or the "official guidelines" that tend to supplant it by confidence intervals. I argue that the sole effective strategy is *a smooth transition towards the Bayesian paradigm* (Lecoutre et al., 2001).

The suggested training strategy is to introduce Bayesian methods as follows: (1) to present natural *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 users with a real possibility of *thinking sensibly about statistical inference* problems and behaving in a more reasonable manner.

3. An illustrative example

My first example of application will concern the inference about a proportion in a clinical trial (Lecoutre et al., 1995). The patients under study were post-myocardial infarction patients, treated with a low-molecular-weight heparin as a prophylaxis of an intra-cardial left ventricular thrombosis. Because of the limited knowledge available on drug potential efficacy, the trial aimed at abandoning further development as early as possible if the drug was likely to be not effective, and at estimating its efficacy if it turned out to be promising. It was considered that 0.85 was the success rate (no thrombosis) above which the drug would be attractive, and that 0.70 was the success rate below which the drug would be of no interest.

The trial was initially designed within the traditional Neyman–Pearson framework. Considering the null hypothesis H_0: $\varphi = 0.70$, the investigators planned a one-sided fixed sample Binomial test with specified respective Type I and Type II error probabilities $\alpha = 0.05$ and $\beta = 0.20$, hence a power $1-\beta = 0.80$ at the alternative H_a: $\varphi = 0.85$ (the hypothesis that they wish to accept!). The associated sample size was $n = 59$, for which the Binomial test rejects H_0 at level 0.05 if the observed number of success a is greater than 47. Indeed, for a sample of size n, the probability of observing a successes is given by the Binomial distribution

$$a|\varphi \sim \mathrm{Bin}(\varphi, n),$$

$$\Pr(a|\varphi) = \binom{n}{a}\varphi^{a}(1-\varphi)^{n-a},$$

hence the likelihood function

$$\ell(\varphi|\text{data}) \sim \varphi^{a}(1-\varphi)^{n-a}.$$

For $n = 59$ (which can be found by successive iterations), we get:

$$\Pr(a > 47 | H_0 : \varphi = 0.70) = 0.035 < 0.05\ (\alpha)$$
$$\Pr(a > 47 | H_\mathrm{a} : \varphi = 0.85) = 0.834 > 0.80\ (1-\beta).$$

Note that, due to the discreteness of the distribution, the actual Type I error rate and the actual power differ from α and $1-\beta$.

Since it would be preferable to stop the experiment as early as possible if the drug was likely to be ineffective, the investigators planned an interim analysis after 20 patients have been included. Since the traditional Neyman–Pearson framework requires specification of all possibilities in advance, they designed a stochastically curtailed test. Stochastic curtailment suggests that an experiment be stopped at an interim stage when the available information determines the outcome of the experiment with high probability under either H_0 or H_a. The notations are summarized in Table 1.

3.1. Stochastically curtailed testing and conditional power

Stochastically curtailed testing uses the "conditional power" at interim analysis, which is defined as the probability, given φ and the available data, that the test rejects H_0 at the planned termination. At interim analysis, termination occurs to reject H_0 if the conditional power at the null hypothesis value is high, say greater than 0.80. In our example, even if after 20 observations 20 successes have been observed, we do not stop the trial.

Similarly, early termination may be allowed to accept H_0 if the conditional power at the alternative hypothesis value is weak, say smaller than 0.20. For instance, if 12 successes have been observed after 20 observations this rule suggests stopping and accepting the null hypothesis. A criticism addressed to this procedure is that there seems little point in considering a prediction that is based on hypotheses that may be no longer fairly plausible given the available data. In fact, the procedure ignores the knowledge about the parameter accumulated by the time of the interim analysis.

Table 1
Summary of the notations for the inference about a proportion

	Number of Successes	Number of Errors	Sample Size
Current data at interim stage	a_1	$n_1 - a_1$	$n_1 = 20$
Future data	a_2	$n_2 - a_2$	$n_1 = 39$
Complete data	$a = a_1 + a_2$	$n - a$	$n = 59$

3.2. An hybrid solution: the predictive power

Many authors have advocated calculating the "predictive power," averaging conditional power over values of the parameter in a Bayesian calculation. We are led to a Bayesian approach, but still with a frequentist test in mind. Formally, the prediction uses the posterior distribution of φ given a prior and the data available at the interim analysis. For the inference about a proportion, the calculations are particularly simple if we choose a conjugate Beta prior distribution

$$\varphi \sim \text{Beta}(a_0, b_0),$$

with density

$$p(\varphi) = \frac{1}{\text{B}(a_0, b_0)} \varphi^{a_0-1}(1-\varphi)^{b_0-1}.$$

The advantage is that the posterior is also a Beta distribution (hence the name conjugate), with density

$$p(\varphi|\text{data}) \propto \ell(\varphi|\text{data}) \times p(\varphi) \propto \varphi^{a_0+a-1}(1-\varphi)^{b_0+b-1}.$$

The prior weights a_0 and b_0 are added to the observed counts a_1 and b_1, so that at the interim analysis

$$\varphi|\text{data} \sim \varphi|a_1 \sim \text{Beta}(a_1 + a_0, b_1 + b_0).$$

The predictive distribution, which is a mixture of Binomial distributions, is naturally called a Beta–Binomial distribution

$$a_2|a_1 \sim \text{Beta}-\text{Bin}(a_1 + a_0, b_1 + b_0; n_2).$$

A vague or *non-informative* prior is generally considered. It is typically defined by small weights a_0 and b_0, included between 0 and 1. Here, I have retained a Beta prior with parameters 0 and 1

$$\varphi \sim \text{Beta}(0, 1).$$

This choice is consistent with the test procedure. I shall address this issue in greater detail later on.

In the example above with $n_1 = 20$ and $a_1 = 20$, the predictive probability of rejecting H_0 at the planned termination ($n = 59$) explicitly takes into account the available data (no failure has been observed). It is with no surprise largely greater than the probability conditional on the null hypothesis value

$$\Pr(a > 47|a_1 = 20) = \Pr(a_2 > 27|a_1 = 20) = 0.997 > 0.80,$$

hence the decision to stop and reject H_0.

This predictive probability is a weighted average of the probabilities conditional to φ, the weights being given by the posterior distribution

$$\Pr(a > 47|a_1 = 20 \text{ and } \varphi) = \Pr(a_2 > 27|a_1 = 20 \text{ and } \varphi),$$

some examples of which being

φ	$\mapsto$	$\Pr(a > 47 \vert a_1 = 20 \text{ and } \varphi)$
1		1
0.95		0.9999997
0.85		0.990
0.70		0.482

Since the predictive power approach is a hybrid one, it is most unsatisfactory. In particular, it does not give us direct Bayesian information about φ. The trouble is that a decision (to accept H_0 or to accept H_a) is taken at the final analysis (or eventually at an interim analysis), even if the observed proportion falls in the no-decision region [0.70, 0.85], in which case *nothing has been proved.*

What the investigators need is to evaluate at any stage of the experiment the probability of some specified regions of interest and the ability of a future sample to support and corroborate findings already obtained. The Bayesian analysis addresses these issues.

3.3. The Bayesian solution

Bayesian methodology enables the probabilities of the pre-specified regions of interest to be obtained. Such statements give straight answers to the question of effect sizes and have no frequentist counterpart. Consider the following example of Bayesian interim analysis, with 10 observed successes ($n_1 = 20$ and $a_1 = 10$).

3.3.1. Evaluating the probability of specified regions

Let us assume the Jeffreys prior Beta(1/2, 1/2) – hence the posterior Beta(10.5, 10.5) shown in Fig. 1 – that will give the privileged non-informative solution (I shall also address this issue later on).

In this case it is very likely that the drug is ineffective ($\varphi < 0.70$), as indicated by the following statements

$$\Pr(\varphi < 0.70 \vert a_1 = 10) = 0.971$$
$$\Pr(0.70 < \varphi < 0.85 \vert a_1 = 10) = 0.029 \quad \Pr(\varphi > 0.85 \vert a_1 = 10) = 0.0001.$$

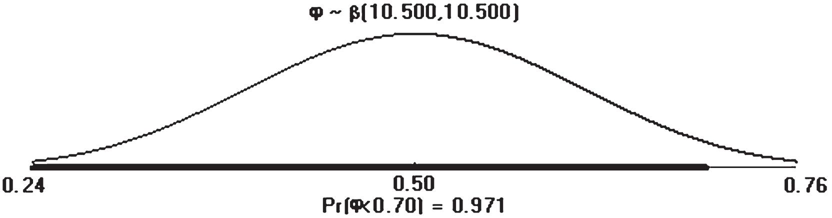

Fig. 1. Example of interim analysis ($n_1 = 20$ and $a_1 = 10$). Density of the posterior distribution Beta(10.5, 10.5) associated with the prior Beta(1/2, 1/2).

Note that in this case, the Bayesian inference about φ at the interim analysis does not explicitly integrate the stopping rule (which is nevertheless taken into account in the predictive probability). In the frequentist framework, the interim inferences are usually modified according to the stopping rule. This issue – that could appear as an area of disagreement between the frequentist and Bayesian approaches – will be considered later on. Resorting to computers solves the technical problems involved in the use of Bayesian distributions. This gives the users 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.

3.3.2. Evaluating the ability of a future sample to corroborate the available results

As a summary to help in the decision whether to continue or to terminate the trial, it is useful to assess the predictive probability of confirming the conclusion of ineffectiveness. If a guarantee of at least 0.95 for the final conclusion is wanted, that is $\Pr(\varphi < 0.70 \mid a) > 0.95$, the total number of successes a must be less than 36 out of 59. Since $a_1 = 10$ successes have been obtained, we must compute the predictive probability of observing $0 \leq a_2 \leq 25$ successes in the future data. Here, given the current data, there is about 87% chance that the conclusion of ineffectiveness will be confirmed. Table 2 gives a summary of the analyses for the previous example and for another example more favorable to the new drug.

3.3.3. Determining the sample size

Predictive procedures are also useful tools to help in the choice of the sample size. Suppose that in order to plan a trial to demonstrate the effectiveness of the drug, we have realized a pilot study: for instance, with $n_0 = 10$ patients, we have observed zero failure. In this case, the posterior probability from the pilot

Table 2
Summary of the Bayesian interim analyses

	Prior Distribution Beta(1/2, 1/2)	
	Example 1: $n_1 = 20$ and $a_1 = 10$	
Inference about φ		Predictive probability ($n = 59$)
Posterior probability		Conclusion with guarantee ≥ 0.95
$\Pr(\varphi < 0.70 \mid a_1 = 10)$		$\varphi < 0.70$
0.971		0.873 ($a < 36$)
$\Pr(\varphi < 0.85 \mid a_1 = 10)$		$\varphi < 0.85$
0.9999		0.9998 ($a < 46$)
	Example 2: $n_1 = 20$ and $a_1 = 18$	
Inference about φ		Predictive probability ($n = 59$)
Posterior probability		Conclusion with guarantee ≥ 0.95
$\Pr(\varphi < 0.70 \mid a_1 = 10)$		$\varphi > 0.70$
0.982		0.939 ($a > 46$)
$\Pr(\varphi < 0.85 \mid a_1 = 10)$		$\varphi > 0.85$
0.717		0.301 ($a > 54$)

experiment (starting with the Jeffreys prior) is used as prior distribution. Here, for this prior, $\Pr(\varphi > 0.85) = 0.932$. If the preliminary data of the pilot study are integrated in the analysis ("full Bayesian" approach), the procedure is exactly the same as that of the interim analysis. However, in most experimental devices, the preliminary data are not included, and the analysis is conducted using a non-informative prior, here Beta(1/2, 1/2).

The procedure remains analogous: we compute the predictive probability that in the future sample of size n (not in the whole data), the conclusion of effectiveness ($\varphi > 0.85$) will be reached with a given guarantee γ. Hence, for instance, the following predictive probabilities for $\gamma = 0.95$

$$\begin{array}{ll} n = 20 \mapsto 0.582\ (a > 19) & n = 30 \mapsto 0.696\ (a > 28) \\ n = 40 \mapsto 0.744\ (a > 37) & n = 50 \mapsto 0.770\ (a > 46) \\ n = 60 \mapsto 0.787\ (a > 55) & n = 70 \mapsto 0.696\ (a > 64) \\ n = 71 \mapsto 0.795\ (a > 65) & n = 72 \mapsto 0.829\ (a > 65). \end{array}$$

Values within parentheses indicate those values of a that satisfy the condition

$$\Pr(\varphi > 0.85 | a) \geq 0.95.$$

Based on the preliminary data, there are 80% chances to demonstrate effectiveness with a sample size about 70. Note that it is not surprising that the probabilities can be non-increasing: this results in the discreteness of the variable (it is the same for power).

3.4. A comment about the choice of the prior distribution: Bayesian procedures are no more arbitrary than frequentist ones

Many potential users of Bayesian methods continue to think that they are too subjective to be scientifically acceptable. However, frequentist methods are full of more or less ad hoc conventions. Thus, the p-value is traditionally based on the samples that are "more extreme" than the observed data (under the null hypothesis). But, for discrete data, it depends on whether the observed data are included or not in the critical region. So, for the usual Binomial one-tailed test for the null hypothesis, $\varphi = \varphi_0$ against the alternative $\varphi > \varphi_0$, this test is *conservative*, but if the observed data are excluded, it becomes *liberal*. A typical solution to overcome this problem consists in considering a mid-p-value, but it has only ad hoc justifications.

In our example, suppose that 47 successes are observed at the final analysis ($n = 59$ and $a = 47$), that is the value above which the Binomial test rejects $H_0{:}\varphi = 0.70$. The p-value can then be computed according to the three following possibilities:

(1) $p_{\text{inc}} = \Pr(a \geq 47 | H_0{:}\ \varphi = 0.70) = 0.066$ ["including" solution]
$\Rightarrow H_0$ is not rejected at level $\alpha = 0.05$ (conservative test)
(2) $p_{\text{exc}} = \Pr(a > 47 | H_0{:}\ \varphi = 0.70) = 0.035$ ["excluding" solution]
$\Rightarrow H_0$ is rejected at level $\alpha = 0.05$ (liberal test)
(3) $p_{\text{mid}} = 1/2(p_{\text{inc}} + p_{\text{exc}}) = 0.051$ [mid-p-value]

Obviously, in this case the choice of a non-informative prior distribution cannot avoid conventions. But the particular choice of such a prior is an exact counterpart of the arbitrariness involved within the frequentist approach. For Binomial sampling, different non-informative priors have been proposed (for a discussion, see, e.g., Lee, 2004, pp. 79–81). In fact, there exist two extreme non-informative priors that are, respectively, the more unfavorable and the more favorable priors with respect to the null hypothesis. They are respectively the Beta distribution of parameters 1 and 0 and the Beta distribution of parameters 0 and 1. These priors lead to the Bayesian interpretation of the Binomial test: the observed significance levels of the inclusive and exclusive conventions are exactly the posterior Bayesian probabilities that φ is greater than φ_0, respectively, associated with these two extreme priors. Note that these two priors constitute an a priori "ignorance zone" (Bernard, 1996), which is related to the notion of imprecise probability (see Walley, 1996).

(1) $\Pr(\varphi < 0.70 | a = 47) = 0.066 = p_{\text{inc}}$
for the prior $\varphi \sim$Beta(0, 1) (the most unfavorable to H_0)
hence the posterior $\varphi|\text{a} \sim$ Beta(47, 13)
(2) $\Pr(\varphi < 0.70 | a = 47) = 0.035 = p_{\text{exc}}$
for the prior $\varphi \sim$ Beta(1, 0) (the most favorable to H_0)
hence the posterior $\varphi|\text{a} \sim$ Beta(48, 12)
(3) $\Pr(\varphi \gtrless 0.70 | a = 47) = 0.049 \approx p_{\text{mid}}$
for the prior $\varphi \sim$Beta(1/2, 1/2)
hence the posterior $\varphi|\text{a} \sim$ Beta(47.5, 12.5)

Then the usual criticism of frequentists towards the divergence of Bayesians with respect to the choice of a non-informative prior can be easily reversed. Furthermore, the Jeffreys prior, which is very naturally the intermediate Beta distribution of parameters 1/2 and 1/2, gives a posterior probability, fully justified, close to the observed mid-*p*-value. The Bayesian response should not be to underestimate the impact of the choice of a particular non-informative prior, as it is often done,

> In fact, the [different non informative priors] do not differ enough to make much difference with even a fairly small amount of data. (Lee, 2004, p. 81)

but on the contrary to assume it.

3.5. Bayesian credible intervals and frequentist coverage probabilities

In other situations, where there is no particular value of interest for the proportion, we may consider an interval (or more generally a region) estimate for φ. In the Bayesian framework, such an interval is usually termed a *credible interval* (or *credibility interval*), which explicitly accounts for the difference in interpretation with the frequentist confidence interval.

3.5.1. Equal-tails intervals

Table 3 gives 95% equal-tails credible intervals for the following two examples, assuming different non-informative priors.

The prior Beta(1, 0), which gives the largest limits, has the following frequentist properties: the proportion of samples for which the upper limit is less than φ is smaller than $\alpha/2$ and the proportion of samples for which the lower limit is more than φ is larger than $\alpha/2$. The prior Beta(0, 1), which gives the smallest limits, has the reverse properties. Consequently, simultaneously considering the limits of these two intervals protects the user both from erroneous acceptation and rejection of hypotheses about φ. This is undoubtedly an objective Bayesian analysis. If a single limit is wanted for summarizing and reporting results, these properties lead to retain the *intermediate* symmetrical prior Beta(1/2, 1/2) (which is the Jeffreys prior). Actually, the Jeffreys credible interval has remarkable frequentist properties. Its coverage probability is very close to the nominal level, even for small-size samples, and it can be favorably compared to most frequentist intervals (Brown et al., 2001; Agresti and Min, 2005).

> We revisit the problem of interval estimation of a Binomial proportion … We begin by showing that the chaotic coverage properties of the Wald interval are far more persistent than is appreciated … We recommend the Wilson interval or the equal-tailed Jeffreys prior interval for small *n*. (Brown et al., 2001, p. 101)

Note that similar results are obtained for *negative*-Binomial (or *Pascal*) sampling, in which we observe the number of patients *n* until a *fixed number of successes a* is obtained. In this case, the observed significance levels of the inclusive and exclusive conventions are exactly the posterior Bayesian probabilities associated with the two respective priors Beta(0, 0) and Beta(0, 1). This suggests privileging the intermediate Beta distribution of parameters 0 and 1/2, which is precisely the Jeffreys prior. This result concerns an important issue related to the "likelihood principle." I shall address it in greater detail further on.

3.5.2. Highest posterior density intervals

A frequently recommended alternative approach is to consider the *highest posterior density* (HPD) credible interval. For such an interval, which can be in fact an union of disjoint intervals (if the distribution is not unimodal), every point included has higher probability density than every point excluded. The aim is to

Table 3
Example of 95% credible intervals assuming different non-informative priors

Beta(0, 1)	Beta(1, 1)	Beta(1/2, 1/2)	Beta(0, 0)	Beta(1, 0)
$n_1 = 20$, $a_1 = 19$				
[0.7513, 0.9877]	[0.7618, 0.9883]	[0.7892, 0.9946]	[0.8235, 0.9987]	[0.8316, 0.9987]
$n_1 = 59$, $a_1 = 32$				
[0.4075, 0.6570]	[0.4161, 0.6633]	[0.4158, 0.6649]	[0.4240, 0.6728]	[0.4240, 0.6728]

get the shortest possible interval. However, except for a symmetric distribution, each of the two one-sided probabilities is different from $\alpha/2$. This property is generally undesirable in experimental data analysis, since more questions are "one-sided" as in the present example.

Moreover, such an interval is not invariant under transformation (except for a linear transformation), which can be considered with Agresti and Min (2005, p. 3) as "a fatal disadvantage." So, for the data $n = 59$, $a = 32$ and the prior Beta(1/2, 1/2), we get the HPD intervals

$$[0.4167, 0.6658] \text{ for } \varphi \text{ and } [0.7481, 2.1594] \quad \text{for} \quad \frac{\varphi}{1-\varphi},$$

with the one-sided probabilities

$$\Pr(\varphi < 0.4167) = 0.026 \text{ and } \Pr\left(\frac{\varphi}{1-\varphi} < 0.7481\right) = 0.039,$$

$$\Pr(\varphi < 0.6658) = 0.024 \text{ and } \Pr\left(\frac{\varphi}{1-\varphi} < 2.1594\right) = 0.011.$$

It must be emphasized, from this example, that the posterior distribution of $\varphi/(1-\varphi)$ is easily obtained: it is a Fisher–Snedecor F distribution. We find the 95% equal-tails interval [0.712, 1.984].

3.6. The contribution of informative priors

When an objective Bayesian analysis suggests a given conclusion, various prior distributions expressing results from other experiments or subjective opinions from specific, well-informed individuals ("experts"), whether *skeptical* or *convinced* (*enthusiastic*), can be investigated to assess the robustness of conclusions (see, in particular, Spiegelhalter et al., 1994).

The elicitation of a prior distribution from the opinions of "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) and should be used with caution. The following examples are provided to understand how the Bayesian inference combines information, and are not intended to correspond to a realistic situation (in the current situation, no good prior information was available). I leave the reader the task to appreciate the potential contribution of these methods.

3.6.1. Skeptical and convinced priors

Consider again the example of data $n = 59$, $a = 32$, for which the objective Bayesian procedure concludes to inefficiency ($\varphi < 0.70$). For the purpose of illustration, let us assume the two priors, a priori, respectively, very skeptical and very convinced about the drug:

$$\varphi \sim \text{Beta}(20, 80) \quad \text{with mean } 0.200 \quad \text{for which } \Pr(\varphi < 0.70) \approx 1,$$
$$\varphi \sim \text{Beta}(98, 2) \quad \text{with mean } 0.980 \quad \text{for which } \Pr(\varphi > 0.85) = 0.999998.$$

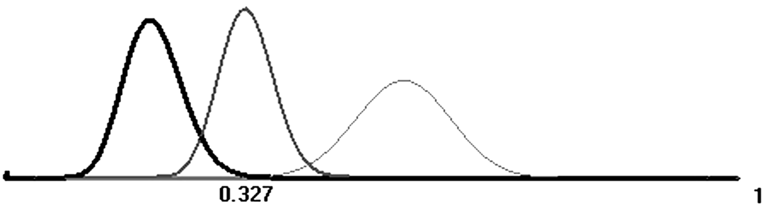

Fig. 2. Example of skeptical prior for the data $n = 59$ and $a = 32$. Densities of the prior Beta(20, 80) (thick line) and of the posterior distributions associated with this prior (medium line) and with the prior Beta(1/2, 1/2) (thin line).

The respective posteriors are

$$\varphi \sim \text{Beta}(52, 107) \quad \text{with mean } 0.327 \quad \text{for which } \Pr(\varphi < 0.70) \approx 1,$$
$$\varphi \sim \text{Beta}(130, 29) \quad \text{with mean } 0.818 \quad \text{for which } \Pr(\varphi > 0.85) = 0.143.$$

Of course the first prior reinforces the conclusion of inefficiency. Figure 2 shows this prior density (thick line) and the posterior (medium line), which can be compared to the objective posterior for the prior Beta(1/2, 1/2) (thin line). However, for the planned sample size, this prior opinion does not have any chance of being infirmed by the data. Even if 59 successes and 0 error had been observed, one would have $\Pr(\varphi < 0.70)|a = 59) = 0.99999995$.

The second prior allows a clearly less unfavorable conclusion. However, the efficiency of the drug cannot be asserted:

$$\Pr(\varphi > 0.70|a = 32) = 0.997 \quad \text{but} \quad \Pr(\varphi > 0.85|a = 32) = 0.143.$$

It is enlightening to examine the impact of the prior Beta(a_0, b_0) on the posterior mean. Letting $n_0 = a_0 + b_0$, the ratios $n_0/(n_0 + n)$ and $n/(n_0 + n)$ represent the relative weights of the prior and of the data. The posterior mean can be written

$$\frac{a_0 + a}{n_0 + n} = \frac{n_0}{n_0 + n}\frac{a_0}{n_0} + \frac{n}{n_0 + n}\frac{a}{n},$$

and is consequently equal to

$$\text{prior relative weight} \times \text{prior mean} + \text{data relative weight} \times \text{observed mean}.$$

The posterior means are as follows:

$$100/159 \times 0.200 + 59/159 \times 0.542 = 0.327 \text{ for the prior } \varphi \sim \text{Beta}(20, 80),$$

$$100/159 \times 0.980 + 59/159 \times 0.542 = 0.818 \text{ for the prior } \varphi \sim \text{Beta}(98, 2).$$

3.6.2. Mixtures of Beta densities

A technique that remains simple to manage is to use a prior with a density defined as a mixture of prior densities of Beta distributions. The posterior is again such a mixture. This prior has two main interests, on the one hand to approximate any arbitrary complex prior that otherwise would need numerical integration methods, and on the other hand to combine several pieces of information (or different opinions). As an illustration, let us consider for the same data a mixture of the two previous distributions with equal weights, that is

$$\varphi \sim \frac{1}{2}\mathrm{Beta}(20, 80) \oplus \frac{1}{2}\mathrm{Beta}(98, 2),$$

where $\oplus$ refers to a mixture of densities, that is symbolically written

$$p(\varphi) = \frac{1}{2}p(\mathrm{Beta}(20, 80)) + \frac{1}{2}p(\mathrm{Beta}(98, 2)).$$

Note that this distribution must not be confounded with the distribution of the linear combination of two variables with independent Beta distributions (that would have a much more complex density).

Figure 3 shows the prior density (thick line), which is bimodal, the corresponding posterior (medium line) and the Jeffreys posterior (thin line). In fact, in this case, the data $n = 59$, $a = 32$ allow us, in some sense, to discriminate between the two distributions of the mixture, as the posterior distribution is

$$0.999999903\mathrm{Beta}(52, 107) \oplus 0.000000097\mathrm{Beta}(130, 29),$$

so that it is virtually confounded with the distribution Beta(52, 107) associated with the prior Beta(20, 80).

It is enlightening to note that the weight associated with each Beta distribution of the posterior mixture is proportional to the product of the prior weight times the predictive probability of the data associated with the corresponding Beta prior.

If the number of patients is multiplied by 10, with the same proportion of successes ($n = 590$, $a = 320$), the posterior density, shown in Fig. 4, is virtually confounded with the posterior Beta(340, 350) associated with the prior Beta(20, 80). Of course, it is closer to the Jeffreys solution.

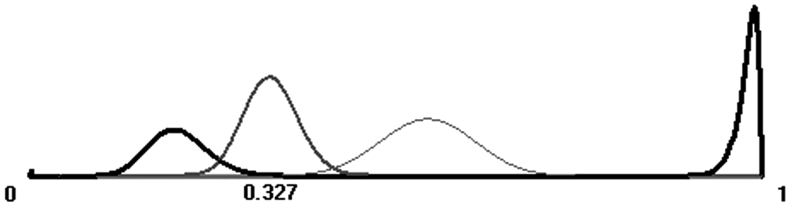

Fig. 3. Example of mixture prior for the data $n = 59$ and $a = 32$. Densities of the bimodal prior (1/2)Beta(20, 80)$\oplus$(1/2)Beta(98, 2) (thick line) and of the posterior distributions associated with this prior (medium line) and with the prior Beta(1/2, 1/2) (thin line).

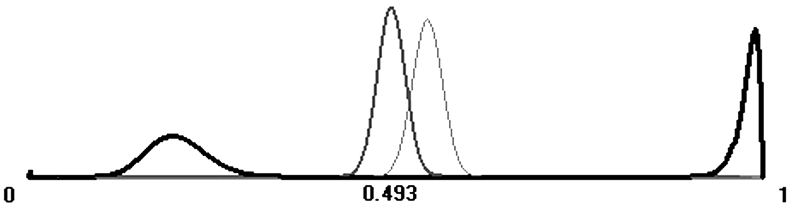

Fig. 4. Example of mixture prior for the data $n = 590$ and $a = 320$. Densities of the bimodal prior (1/2)Beta(20, 80)⊕(1/2)Beta(98, 2) (thick line) and of the posterior distributions associated with this prior (medium line) and with the prior Beta(1/2, 1/2) (thin line).

3.7. The Bayes factor

In order to complete the presentation of the Bayesian tools, I shall present the *Bayes factor*. Consider again the example of data $n = 59$, $a = 32$, with the convinced prior $\varphi \sim \text{Beta}(98, 2)$ and the corresponding a priori probabilities $\Pr(\varphi > 0.85) = 0.99999810$ (that will be denoted π_a), and consequently $\Pr(\varphi < 0.85) = 0.00000190$ (π_0). The notations π_0 and π_a are usual, since the Bayes factor is generally presented as a Bayesian approach to classical hypothesis testing; in this framework, π_0 and π_a are the respective prior probabilities of the null H_0 and alternative H_a hypotheses.

It is then quite natural to consider:

- the ratio of these two prior probabilities, hence

$$\frac{\pi_0}{\pi_\text{a}} = \frac{\Pr(\varphi < 0.85)}{\Pr(\varphi > 0.85)} = 0.0000019,$$

which here is of course very small,

- and their posterior ratio, hence

$$\frac{p_0}{p_\text{a}} = \frac{\Pr(\varphi < 0.85|a = 32)}{\Pr(\varphi > 0.85|a = 32)} = \frac{0.8570}{0.1430} = 5.99,$$

which is now distinctly larger than 1.

The Bayes factor (associated with the observation a) is then defined as the ratio of these two ratios

$$B(a) = \frac{p_0/p_\text{a}}{\pi_0/\pi_\text{a}} = \frac{p_0\pi_\text{a}}{p_\text{a}\pi_0} = 3154986,$$

which evaluates the modification of the relative likelihood of the null hypothesis due to the observation. However, the Bayes factor is only an incomplete summary, which cannot replace the information given by the posterior probabilities.

The Bayes factor applies in the same way to non-complementary hypotheses H_0 and H_a, for instance, here $\varphi < 0.70$ and $\varphi > 0.85$. However, in this case the interpretation is again more problematic, since the "no-decision" region $0.70<\varphi<0.85$ is ignored.

In the particular case of two simple hypotheses H_0: $\varphi = \varphi_o$ and H_a: $\varphi = \varphi_a$, the Bayes factor is simply the classical *likelihood ratio*

$$B(a) = \frac{p(\varphi_0|a)p(\varphi_a)}{p(\varphi_a|a)p(\varphi_0)} = \frac{p(a|\varphi_0)}{p(a|\varphi_a)},$$

since $p(\varphi_0|a) \propto p(a|\varphi_0)p(\varphi_0)$ and $p(\varphi_a|a) \propto p(a|\varphi_a)p(\varphi_a)$.

Note again that when H_0 and H_a are complementary hypotheses (hence $p_a = 1–p_0$), as in the example above, their posterior probabilities can be computed from the prior probabilities ($\pi_a = 1-\pi_0$) and the Bayes ratio. Indeed, it can be easily verified that

$$\frac{1}{p_0} = 1 + \frac{1-\pi_0}{\pi_0}\frac{1}{B(a)}.$$

4. Other examples of inferences about proportions

4.1. Comparison of two independent proportions

Conceptually, all the Bayesian procedures for a proportion can be easily extended to two Binomial independent samples, assuming two independent priors (see Lecoutre et al., 1995). In order to illustrate the conceptual simplicity and the flexibility of Bayesian inference, I give in the subsequent subsection an application of these procedures for a different sampling model.

4.2. Comparison of two proportions for the play-the-winner rule

From ethical point of view, adaptative designs can be desirable. In such designs subjects are assumed to arrive sequentially and they are assigned to a treatment with a probability that is updated as a function of the previous events. The intent is to favor the "most effective treatment" given available information. The *play-the-winner* allocation rule is designed for two treatments t^1 and t^2 with a dichotomous (e.g., success/failure) outcome (Zelen, 1969). It involves an "all-or-none" process: if subject $k-1$ is assigned to treatment t (t^1 or t^2) and if the outcome is a success (with probability φ_t), subject k is assigned to the same treatment; if, on the contrary, the outcome is a failure (with probability $1-\varphi_t$), subject k is assigned to the other treatment.

For simplicity, it is assumed here that the outcome of subject $k-1$ is known when subject k is included.

For a fixed number n of subjects, the sequel of treatment allocations (t_1, t_2,..., t_k, t_{k+1}, ..., t_{n+1}) contains all the information in the data. Indeed, $t_k = t_{k+1}$ implies that a success to t_k has been observed and $t_k \neq \mathrm{t}_{k+1}$ implies that a failure to t_k has been observed. Moreover, the likelihood function is simply

$$\ell(\varphi_1, \varphi_2)|(t_1, \ldots, t_{n+1}) = \frac{1}{2}\varphi_1^{n_{11}}(1-\varphi_1)^{n_{10}}\varphi_2^{n_{21}}(1-\varphi_2)^{n_{20}},$$

where n_{ij} is the number of pairs (t_k, t_{k+1}) equal to (t^i, t^j), so that n_{11} and n_{21} are the respective numbers of success to treatments t^1 and t^2, and n_{10} and n_{20} are the numbers of failure (1/2 is the probability of t_1).

Since Bayesian methods only involve the likelihood function, they are immediately available. Moreover, since the likelihood function is identical (up to a multiplicative constant) with the likelihood function associated with the comparison of two independent binomial proportions, the same Bayesian procedures apply here, even if the sampling probabilities are very different. On the contrary, with the frequentist approach, specific procedures must be developed. Due to the complexity of the sampling distribution, only asymptotic solutions are easily available. Of course, except for large samples, they are not satisfactory.

4.2.1. Numerical example

Let us consider for illustration the results of a trial with $n = 150$ subjects. The observed rates of success are, respectively, 74 out of 94 attributions for treatment t^1 and 35 out of 56 attributions for treatment t^2. Note that, from the definition of the rule, the numbers of failures (here 20 and 21) can differ at most by 1. A joint probability statement is, in a way, the best summary of the posterior distribution. For instance, if we assume the Jeffreys prior, that is two independent Beta(1/2, 1/2) distributions for φ_1 and φ_2, the marginal posteriors Beta(74.5, 20.5) and Beta(35.5, 21.5) are again independent, so that a joint probability statement can be immediately obtained. We get, for instance,

$$\Pr(\varphi_1 > 0.697 \text{ and } \varphi_2 < 0.743|\text{data}) = 0.95$$

which is deduced from $\Pr(\varphi_1 > 0.697) = \Pr(\varphi_2 > 0.743) = \sqrt{0.95} = 0.9747$, obtained as in the case of the inference about a single proportion.

It is, in a way, the best summary of the posterior distribution. However, a statement that deals with the comparison of the two treatments directly would be preferable. So we have a probability 0.984 that $\varphi_2 > \varphi_1$. Furthermore, the distribution of any derived parameter of interest can be easily obtained from the joint posterior distribution using numerical methods. We find the 95% equal-tails credible intervals:

$$[+0.013, +0.312] \text{ for } \varphi_1 - \varphi_2[1.02, 1.62] \text{ for } \frac{\varphi_1}{\varphi_2}[1.07, 4.64] \text{ for } \frac{\varphi_1/(1-\varphi_1)}{\varphi_2/(1-\varphi_2)}.$$

For the Jeffreys prior, Bayesian methods have fairly good frequentist coverage properties for interval estimates (Lecoutre and ElQasyr, 2005).

4.2.2. The reference prior approach

For multidimensional parameter problems, the *reference prior* approach introduced by Bernardo (1979) (see also Berger and Bernardo, 1992) can constitute a successful refinement of the Jeffreys prior. This approach presupposes that we are interested in a particular derived parameter θ. It aims at finding the optimal objective prior, given that θ is the parameter of interest and the resulting prior is consequently dependent on this parameter. An objection can be raised against this approach in the context of experimental data analysis. Even when a particular parameter is privileged to summarize the findings, we are also interested in other parameters, so that joint prior and posterior distributions are generally wanted.

4.3. A generalization with three proportions: medical diagnosis

Berger (2004, p. 5) considered the following situation (Mossman and Berger, 2001; see also in a different context Zaykin et al., 2004).

> Within a population for which $\varphi_0 = \Pr(\text{Disease } D)$, a diagnostic test results in either a Positive $(+)$ or Negative $(-)$ reading. Let $\varphi_1 = \Pr(+|\text{patient has } D)$ and $(\varphi_2 = \Pr(+|\text{patient does not have } D)$. [the authors notations p_i have been changed to φ_i]

By Bayes' theorem, one get the probability θ that the patient has the disease given a positive diagnostic test

$$\theta = \Pr(D|+) = \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + \Pr(+|-D)\Pr(-D)} = \frac{\varphi_1\varphi_0}{\varphi_1\varphi_0 + \varphi_2(1-\varphi_0)}.$$

It is assumed that for $i = 0, 1, 2$ there are available (independent) data a_i, having Binomial distributions

$$a_i|\varphi_i \sim \text{Bin}(\varphi_i, n_i),$$

hence a straightforward generalization of the inference about two independent proportions. Note that, conditionally to φ_0, the situation is that of inference about the ratio of two independent Binomial proportions, since for instance

$$\Pr(\theta < u|\varphi_0) = \Pr\left(\frac{\varphi_2}{\varphi_1} > \frac{1-\varphi_0}{\varphi_0}\frac{1-u}{u}\right).$$

The marginal probability is a mixture of these conditional probabilities.

It results "a simple and easy to use procedure, routinely usable on a host of applications," which, from a frequentist perspective "has better performance [...] than any of the classically derived confidence intervals" (Berger, 2004, pp. 6–7).

Another situation that involves a different sampling model but leads to the same structure is presented in greater detail hereafter.

4.4. Logical models in a contingency table

Let us consider a group of n patients, with two sets of binary attributes, respectively, $V = \{v1, v0\}$ and $W = \{w1, w0\}$. To fix ideas, let us suppose that W is cardiac mortality (yes/no) and that V is myocardial infarction (yes/no). Let us consider the following example of logical model (Lecoutre and Charron, 2000).

> An absolute (or logical) *implication* $v1 \Rightarrow\ w1$ (for instance) exists if all the patient having the modality v_1 also have the modality $w1$, whereas the converse is not necessarily true.
>
> However, the hypothesis of an absolute implication (here "myocardial infarction implies cardiac mortality") is of little practical interest, since a single observation of the event ($v1$, $w0$) is sufficient to falsify it.
>
> Consequently, we have to consider the weaker hypothesis "$v1$ implies in most cases $w0$" ($v1 \hookrightarrow w1$).

The issue is to evaluate the departure from the logical model "the cell ($v1$, $w0$) should be empty." A departure index $\eta_{v1 \hookrightarrow w1}$ can be defined from the cell proportions

	$W1$	$w0$	
$v1$	φ_{11}	φ_{10}	$\varphi_{1.}$
$v0$	φ_{01}	φ_{00}	$\varphi_{0.}$
	$\varphi_{.1}$	$\varphi_{.0}$	1

as

$$\eta_{v1 \hookrightarrow w1} = 1 - \frac{\varphi_{10}}{\varphi_{1.}\varphi_{.0}} \quad (-\infty < \eta_{v1 \hookrightarrow w1} < +1).$$

This index has been actually considered in various frameworks, with different approaches. It can be viewed as a measure of *predictive efficiency* of the model when predicting the outcome of W given $v1$.

- The prediction is perfect (there is an absolute implication) when $\eta_{v1 \hookrightarrow w1} = +1$.
- The closer to 1 $\eta_{v1 \hookrightarrow w1}$ is, the more efficient the prediction.
- In case of independence, $\eta_{v1 \hookrightarrow w1} = 0$.
- A null or negative value means that the model is a prediction failure.

Consequently, in order to investigate the predictive efficiency of the model, we have to demonstrate that $\eta_{v1 \hookrightarrow w1}$ has a value close to $+1$. Of course, one can define in the same way the indexes $\eta_{v1 \hookrightarrow w0}$, $\eta_{w1 \hookrightarrow v1}$, and $\eta_{w0 \hookrightarrow v0}$ One can, again, characterize the *equivalence* between two modalities. An absolute equivalence between $v1$ and $w1$ (for instance) exists if $\eta_{v1 \hookrightarrow w1} = +1$ and $\eta_{v0 \hookrightarrow w0} = +1$ (the two cells [$v1$, $w0$] and [$v0$, $w1$] are empty). Consequently, the minimum of these two indexes is an index of departure from equivalence.

Let us assume a multinomial sampling model, hence for a sample of size n, the probability of observing the cell counts n_{ij}

$$\Pr(n_{11}, n_{10}, n_{01}, n_{00} | \varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00}) = \frac{n!}{n_{11}!n_{10}!n_{01}!n_{00}!} \varphi_{11}^{n_{11}} \varphi_{10}^{n_{10}} \varphi_{01}^{n_{01}} \varphi_{00}^{n_{00}}.$$

4.5. Frequentist solutions

Asymptotic procedures (see, e.g., Fleiss, 1981) are clearly inappropriate for small samples. Alternative procedures based on Fisher's conditional test (Copas and Loeber, 1990; Lecoutre and Charron, 2000) have been proposed. This test involves the sampling distribution of n_{11} (for instance). A classical result is that this distribution, given fixed observed margins, only depends on the cross product $\rho = \varphi_{11}\varphi_{00}/\varphi_{10}\varphi_{01}$ (Cox, 1970, p. 4). The null hypothesis $\rho = \rho_0$ can be tested against the alternative $\rho < \rho_0$ (or against $\rho > \rho_0$), by using the probability that n_{11} exceeds the observed value in the appropriate direction.

Consequently, the procedure is analogous to the Binomial test considered for the inference about a proportion. We can define in the same way an "including" solution and an "excluding" solution.

In the particular case $\rho_0 = 0$, this test is the Fisher's randomization test of the null hypothesis $\rho = 1$ (i.e., $\eta_{v1 \hookrightarrow w1} = 0$) against $\rho < 1$ ($\eta_{v1 \hookrightarrow w1} < 0$).

By inverting this conditional test, confidence intervals can be computed for the cross product ρ. An interval for $\eta_{v1 \hookrightarrow w1}$ is then deduced by replacing ρ by its confidence limits in the following expression that gives $\eta_{v1 \hookrightarrow w1}$ as a function of ρ

$$\eta_{v1 \hookrightarrow w1} = \frac{1 + (\rho - 1)(\varphi_{1.} + \varphi_{.1} - \varphi_{1.}\varphi_{.1} - [(1 + (\varphi_{1.} + \varphi_{.1})(\rho - 1)^2 - 4\varphi_{1.}\varphi_{.1}\rho(\rho - 1)]^{1/2}}{2(\rho - 1)\varphi_{.1}(1 - \varphi_{1.})}.$$

Unfortunately, these limits depend on the true margin values $\varphi_{.1}$ and $\varphi_{1.}$. The most common procedure consists in simply replacing these *nuisance parameters* by their estimates $f_{.1}$ and $f_{1.}$. It is much more performing than asymptotic solutions, but is unsatisfactory for extreme parameter values. More efficient principles for dealing with nuisance parameters exist (for instance, Toecher, 1950; Rice, 1988). However, one comes up against a problem that is eternal within the frequentist inference, and that is of course entirely avoided in the Bayesian approach. In any case, Bayesian inference copes with the problem of nuisance parameters. Moreover, it explicitly handles the problems of discreteness and unobserved events (null counts) by way of the prior distribution.

4.6. The Bayesian solution

The Bayesian solution is a direct generalization of the Binomial case. Let us assume a joint (conjugate) *Dirichlet* prior distribution, which is a multidimensional extension of the Beta distribution

$$(\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00}) \sim \text{Dirichlet}(\nu_{11}, \nu_{10}, \nu_{01}, \nu_{00}).$$

The posterior distribution is also a Dirichlet in which the prior weights are simply added to the observed cell counts.

$$(\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00})|\text{data} \sim \text{Dirichlet}(n_{11} + \nu_{11}, n_{10} + \nu_{10}, n_{01} + \nu_{01}, n_{00} + \nu_{00}).$$

From the basic properties of the Dirichlet distribution (see, e.g., Bernardo and Smith, 1994, p. 135), the marginal posterior distribution for the derived parameter

η_{11} can be characterized as a function of three independent Beta distributions

$$\begin{aligned}
X &= \varphi_{10}|\text{data} \sim \text{Beta}(n_{10}+\nu_{10}, n_{11}+\nu_{11}+\eta_{01}+\upsilon_{01}+\eta_{00}+\upsilon_{00}),\\
Y &= \frac{\varphi_{00}}{1-\varphi_{10}} = \frac{\varphi_{00}}{1-X}|\text{data} \sim \text{Beta}(n_{00}+\upsilon_{00}, n_{11}+\nu_{11}, n_{01}+\upsilon_{01}),\\
Z &= \frac{\varphi_{11}}{1-\varphi_{10}-\varphi_{00}} = \frac{\varphi_{11}}{(1-Y)(1-X)}|\text{data} \sim \text{Beta}(n_{11}+\nu_{11}, n_{01}+\upsilon_{01}),
\end{aligned}$$

since

$$\eta_{v1 \hookrightarrow w1} = 1 - \frac{X}{(X + Z(1-Y)(1-X))(X + Y(1-X))}$$

This leads to straightforward numerical methods.

4.7. Numerical example: mortality study

4.7.1. Non-treated patients

The data in Table 4 were obtained for 340 high-risk patients who received no medical treatment. Let us consider the implication "Myocardial infarction ↪ Cardiac mortality within 2 years."

The observed values of the index are

- for the implication "Infarction ↪ Decease" (cell [yes,no] empty): $H_{v1 \hookrightarrow w1} = 0.12$,
- for the implication "Decease ↪ Infarction" (cell [no,yes] empty): $H_{v1 \hookrightarrow w1} = 0.37$.

The marginal proportions of decease are (fortunately!) rather small – respectively, 0.22 after infarction and 0.07 without infarction – so that the count 72 in the cell [yes,no] is proportionally large. Consequently, relatively small values of the index are here "clinically significant." Assuming the Jeffreys prior Dirichlet(1/2, 1/2, 1/2, 1/2), we get the posterior

$$\Phi = (\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00})|\text{data} \sim \text{Dirichlet}(20.5, 72.5, 17.5, 231.5).$$

from which we derive the marginal posteriors. Figure 5 shows the decreasing distribution function of the posterior of $\eta_{v1 \hookrightarrow w1}$ and its associated 90% credible interval.

Table 4
Mortality data for 340 high-risk patients who received no medical treatment

		Decease			
		Yes	No		
Myocardial infarction	Yes	20	72	92	[20/92 = 0.22]
	No	17	231	248	[17/248 = 0.07]
		37	303	340	

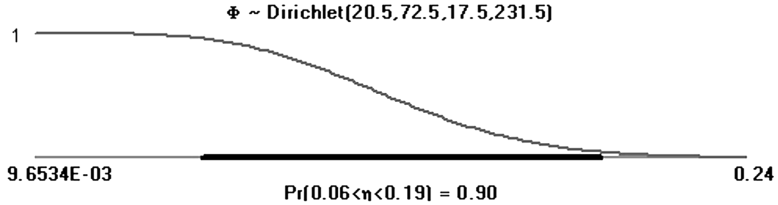

Fig. 5. Implication "Infarction $\hookrightarrow$ Decease" (non-treated patients). Decreasing distribution function for $\eta_{v1\hookrightarrow w1}$ [$\Pr(\eta_{v1\hookrightarrow w1}<x)$] associated with the prior Dirichlet(1/2, 1/2, 1/2, 1/2).

Table 5
Mortality data for 357 high-risk patients who received a preventive treatment

		Decease			
		Yes	No		
Myocardial infarction	Yes	1	78	79	[1/79 = 0.01]
	No	13	265	278	[13/278 = 0.05]
		14	343	357	

From the two credible intervals,

- "Infarction $\hookrightarrow$ Decease": $\Pr(+0.06<\eta_{v1\hookrightarrow w1}<+0.19)=0.90$
- "Decease $\hookrightarrow$ Infarction": $\Pr(+0.20<\eta_{w1\hookrightarrow v1}<+0.54)=0.90$.

we can assert an implication of limited importance. In fact, it appears that decease is a better prognostic factor for infarction than the reverse.

4.7.2. Treated patients

Other data reported in Table 5 were obtained for 357 high-risk patients who received a preventive treatment.

Here, it is, of course, expected that the treatment would reduce the number of deceases after infarction. Ideally, if there was no cardiac decease among the treated patients after infarction (cell [yes,yes] empty), there would be an absolute implication "Infarction $\Rightarrow$ No decease." We get the following results for this implication:

"Infarction $\hookrightarrow$ No decease" : $H_{v1\hookrightarrow w0}=+0.68$ and $\Pr(-0.10<\eta_{v1\hookrightarrow w0}<+0.94)=0.90$.

Here, in spite of a distinctly higher observed value, it cannot be concluded to the existence of an implication. The width of the credible interval shows a poor precision. This is a consequence of the very small observed proportions of decease. Of course, it cannot be concluded that there is no implication or that the implication is small. This illustrate the abuse of interpreting the non-significant result of usual "tests of independence" (chi-square for instance) in favor of the null hypothesis.

4.8. Non-informative priors and interpretation of the observed level of Fisher's permutation tests

The Bayesian interpretation of the permutation test (conditional to margins) generalizes the interpretation of the Binomial test. For the usual one-sided test (including solution), the null hypothesis H_0: $\eta_{v1 \hookrightarrow w0} = 0$ is not rejected ($p_{\text{inc}} = 0.145$). It is well known that this test is conservative, but if we consider the excluding solution, we get a definitely smaller p-value $p_{\text{exc}} = 0.028$. This results from the poor experimental accuracy. As in the case of a single proportion, there exist two extreme non-informative priors, Dirichlet(1, 0, 0, 1) and Dirichlet(0, 1, 1, 0) that constitute the ignorance zone. They give an enlightening interpretation of these two p-values, together with an objective Bayesian analysis.

(1) $\Pr(\eta_{v1 \hookrightarrow w0} < 0) = 0.145 = p_{\text{inc}}$
for the prior Dirichlet(1, 0, 0, 1) (the most unfavorable to H_0)
hence the posterior Dirichlet(2, 78, 13, 266)
(2) $\Pr(\eta_{v1 \hookrightarrow w0} < 0) = 0.028 = p_{\text{exc}}$
for the prior Dirichlet(0, 1, 1, 0) (the most favorable to H_0)
hence the posterior Dirichlet(1, 79, 14, 265)
(3)$\Pr(\eta_{v1 \hookrightarrow w0} < 0) = 0.072 \approx (p_{\text{inc}} + p_{\text{exc}})/2 = 0.086$
for the prior Dirichlet(1/2, 1/2, 1/2, 1/2)
hence the posterior Dirichlet(1.5, 78.5, 13.5, 265.5)

4.8.1. The choice of a non-informative prior

As for a single proportion, the choice of a non-informative prior is no more arbitrary or subjective than the conventions of frequentist procedures. Moreover, simulation studies of frequentist coverage probabilities favorably compare Bayesian credible intervals with conditional confidence intervals (Lecoutre and Charron, 2000). For each lower and upper limits of the $1-\alpha$ credible interval, the frequentist error rates associated with the two *extreme* priors always include $\alpha/2$. Moreover, if a single limit is wanted for summarizing and reporting results, the symmetrical *intermediate* prior Dirichlet(1/2, 1/2, 1/2, 1/2) has fairly good coverage properties, including the cases of moderate sample sizes and small parameter values. Of course the differences between the different priors in the ignorance zone is less for small or medium values of $\eta_{v1 \hookrightarrow w1}$ and vanishes as the sample size increases.

4.9. Further analyses

There is no difficulty in extending the Bayesian procedures to any situation involving the multinomial sampling model, for instance, the comparison of two proportions based on paired data. Here, in particular, the distribution of the minimum of the two indexes for asserting equivalence is easily obtained by simulation. Moreover, the procedures can be extended to compare the indexes associated with two independent groups (for instance, here treated and non-treated patients).

Of course, in all these situations, informative priors and predictive probabilities can be used in the same way as for a single proportion.

Note again that binary and polychotomous response data can also be analyzed by Bayesian regression methods. Relevant references are Albert and Chib (1993) and Congdon (2005).

5. Concluding remarks and some further topics

Time's up to come to a positive agreement for procedures of experimental data analysis that bypass the common misuses of NHST. This agreement should fills up its role of ''an aid to judgment,'' which ''should not be confused with automatic acceptance tests, or 'decision functions''' (Fisher, 1990/1925, p. 128). Undoubtedly, there is an increasing acceptance that Bayesian inference can be ideally suited for this purpose. It fulfills the requirements of scientists: objective procedures (including traditional *p*-values), procedures about effect sizes (beyond *p*-values) and procedures for designing and monitoring experiments. Then, why scientists, and in particular experimental investigators, really appear to want a different kind of inference but seem reluctant to use Bayesian inferential procedures in practice? In a very lucid paper, Winkler (1974, p. 129) 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.'' He concluded that if we leave to one side the choice of philosophical approach, none of the mentioned arguments are entirely convincing. Although Winkler's paper was written more than 30 years ago, it appears as if it had been written today.

> We [statisticians] will all be Bayesians in 2020, and then we can be a united profession. (Lindley, in Smith, 1995, p. 317)

In fact the times we are living in at the moment appear to be crucial. On the one hand, an important practical obstacle is that the standard statistical packages that are nowadays extensively used do not include Bayesian methods. On the other hand, one of the decisive factors could be the recent ''draft guidance document'' of the US Food and Drug Administration (FDA, 2006). This document reviews ''the least burdensome way of addressing the relevant issues related to the use of Bayesian statistics in medical device clinical trials.'' It opens the possibility for experimental investigators to really be Bayesian in practice.

5.1. Some advantages of Bayesian inference

5.1.1. A better understanding of frequentist procedures

> Students [exposed 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. (Berry, 1997)

To take another illustration, let us consider the basic situation of the inference about the difference δ between two normal means. It is especially illustrative of how the Bayesian procedures combine descriptive statistics and significance tests.

Let us denote by d (assuming $d \neq 0$) the observed difference and by t the value of the Student's test statistic. Assuming the usual non-informative prior, the posterior for δ is a generalized (or scaled) t distribution (with the same degrees of freedom as the t-test), centered on d and with scale factor the ratio $e = d/t$ (see, e.g., Lecoutre, 2006a).

From this *technical* link with the t statistic, it results *conceptual* links. The one-sided p-value of the t-test is exactly the posterior Bayesian probability that the difference δ has the opposite sign of the observed difference. Given the data, if for instance $d > 0$, there is a p posterior probability of a negative difference and a $1-p$ complementary probability of a positive difference. In the Bayesian framework these statements are *statistically correct*. Another important feature is the interpretation of the usual confidence interval in natural terms. It becomes correct to say that "there is a 95% [for instance] probability of δ being included between the fixed bounds of the interval" (conditionally on the data).

In this way, Bayesian methods allow users to overcome usual difficulties encountered with the frequentist approach. In particular, using the Bayesian interpretations of significance tests and confidence intervals in the language of probabilities about unknown parameters is quite natural for the users. In return, the common misuses and abuses of NHST are more clearly understood. In particular, users of Bayesian methods become quickly alerted that non-significant results cannot be interpreted as "proof of no effect."

5.1.2. Combining information from several sources

An analysis of experimental data should always include an objective Bayesian analysis in order to express *what the data have to say* independently of any outside information. However, informative Bayesian priors also have an important role to play in experimental investigations. They may help refining inference and investigating the sensitivity of conclusions to the choice of the prior. 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.

Informative Bayesian techniques are ideally suited for *combining information* from the data in hand and from other studies, and therefore planning a series of experiments. More or less realistic and convincing uses have been proposed (for a discussion of how to introduce these techniques in medical trials, see, e.g., Irony and Pennello, 2001). Ideally, when "good prior information is available," it could (should) be used to reach the same conclusion that an "objective Bayesian analysis," but with a smaller sample size. Of course, they should integrate a real knowledge based on data rather than expert opinions, which are generally controversial. However, in my opinion, the use of these techniques must be more extensively explored before appreciating their precise contribution to experimental data analysis.

5.1.3. The predictive probabilities: a very appealing tool

> An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results. (Berry, 1991, p. 81)

A major strength of the Bayesian paradigm is the ease with which one can make predictions about future observations. The predictive idea is central in experimental investigations, as ''the essence of science is replication: a scientist should always be concerned about what would happen if he or another scientist were to repeat his experiment'' (Guttman, 1983). 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 values 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 credible limits give a useful summary to help in the choice of the sample size of an experiment (for parallels between Bayesian and frequentist methods, see Inoue et al., 2005).

The predictive approach is a very appealing method (Baum et al., 1989) to aid the decision to stop an experiment at an interim stage. On the one hand, if the predictive probability that it will be successful appears poor, it can be used as a rule to abandon the experiment for futility. On the other hand, if the predictive probability is sufficiently high, this suggests to early stop the experiment and conclude success.

Predictive probabilities are also a valuable tool for missing data imputation. Note that interim analyses are a kind of such imputation. The case of censored survival data is particularly illustrative. At the time of interim analysis, available data are divided into three categories: (1) included patients for whom the event of interest has been observed, (2) included patients definitely censored and (3) included patients under current observation for whom the maximum observation period has not ended. Consequently, the missing data to be predicted are respectively related to these last patients for which we have partial information and to the new patients planned to be included for which we have no direct information. The Bayesian approach gives us straightforward and effective ways to deal with this situation (Lecoutre et al., 2002).

It can again be outlined that the predictive distributions are also a useful tool for constructing a subjective prior, as it is often easier to express an opinion relative to expected data.

5.2. Bayesian computations and statistical packages

There is currently increasingly widespread application of Bayesian inference for experimental data analysis. However, an obstacle to the routine use of objective Bayesian methods is the lack of user-friendly general purpose software that would be a counterpart to the standard frequentist software. This obstacle may be expected to be removed in the future. Some packages have been designed to learn elementary Bayesian inference: see, for example, First Bayes (O'Hagan, 1996) and a package of Minitab macros (Albert, 1996). With a more ambitious perspective, we have developed a statistical software for Bayesian analysis of variance (Lecoutre and Poitevineau, 1992; Lecoutre, 1996). It incorporates both traditional frequentist practices (significance tests, confidence intervals) and routine 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. This software also includes the basic Bayesian procedures for inference about proportions presented in this chapter.

At a more advanced level, the privileged tool for the Bayesian analysis of complex models is a method called Markov Chain Monte Carlo (MCMC). The principle of MCMC techniques (Gilks et al., 1996; Gamerman, 1997) is to simulate, and consequently approximate, the posterior and predictive distributions (when they cannot be determined analytically). This can be done for virtually any Bayesian analysis. WinBUGS (a part of the BUGS project) is an any general purpose flexible and efficient Bayesian software. It "aims to make practical MCMC methods available to applied statisticians" and largely contributes to the increasing use of Bayesian methods. It can be freely downloaded from the web site: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. However, it can hardly be recommended to beginners unless they are highly motivated.

Very recently, Bayesian analysis has been added in some procedures of the SAS/STAT software. In addition to the full functionality of the original ones, the new procedures produce Bayesian modeling and inference capability in generalized linear models, accelerated life failure models, Cox regression models, and piecewise constant baseline hazard models (SAS Institute Inc., 2006).

5.3. Some further topics

I do not intend to give here an exhaustive selection of topics, but rather to simply outline some areas of research that seems to me particularly important for the methodological development of objective Bayesian analysis for experimental data.

5.3.1. The interplay of frequentist and Bayesian inference

Bayarri and Berger (2004) gave an interesting view of the interplay of frequentist and Bayesian inference. They argued that the traditional frequentist argument, involving "repetitions of the same problem with different data" is not what is done in practice. Consequently, it is "a joint frequentist–Bayesian principle" that is practically relevant: a given procedure (for instance, a 95% confidence interval for a normal mean) is in practice used "on a series of different problems involving a series of different normal means with a corresponding series of data" (p. 60). More generally, they reviewed current issues in the Bayesian–frequentist synthesis from a methodological perspective. It seems a reasonable conclusion to hope a methodological unification, but not a philosophical unification.

> Philosophical unification of the Bayesian and frequentist positions is not likely, nor desirable, since each illuminates a different aspect of statistical inference. We can hope, however, that we will eventually have a general methodological unification, with both Bayesians and frequentists agreeing on a body of standard statistical procedures for general use. (Bayarri and Berger, 2004, p. 78)

In this perspective, an active area of research aims at finding "probability matching priors" for which the posterior probabilities of certain specified sets are equal (at least approximately) to their coverage probabilities: see Fraser et al. (2003) and Sweeting (2005).

5.3.2. Exchangeability and hierarchical models

Roughly speaking, random events are *exchangeable* "if we attribute the same probability to an assertion about any given number of them" (de Finetti, 1972, p. 213). This is a key notion in statistical inference. For instance, future patients must be assumed to be exchangeable with the patients who have already been observed in order to make predictive probabilities reasonable. In the same way, similar experiments must be assumed to be exchangeable for a coherent integration of the information.

The notion of exchangeability is very important and useful in the Bayesian framework. Using multilevel prior specifications, it allows a flexible modeling of related experimental devices by means of *hierarchical models* (Bernardo, 1996).

> If a sequence of observations is judged to be exchangeable, then any subset of them must be regarded as a random sample from some model, and there exist a prior distribution on the parameter of such model, hence requiring a Bayesian approach. (Bernardo, 1996, p. 5)

Hierarchical models are important to make full use of the data from a multicenter experiment. They are also particularly suitable for meta-analysis in which we have data from a number of relevant studies that may be exchangeable on some levels but not on others (Dumouchel, 1990). In all cases, the problem can be decomposed into a series of simpler conditional models, using the hierarchical Bayesian methodology (Good, 1980).

5.3.3. The stopping rule principle: a need to rethink

Experimental designs often involve interim looks at the data for the purpose of possibly stopping the experiment before its planned termination. Most experimental investigators feel that the possibility of early stopping cannot be ignored, since it may induce a bias on the inference that must be explicitly corrected.

Consequently, they regret the fact that the Bayesian methods, unlike the frequentist practice, generally ignore this specificity of the design. Bayarri and Berger (2004) considered this desideratum as an area of current disagreement between the frequentist and Bayesian approaches. This is due to the compliance of most Bayesians with the *likelihood principle* (a consequence of Bayes' theorem), which implies the *stopping rule principle* in interim analysis:

> Once the data have been obtained, the reasons for stopping experimentation should have no bearing on the evidence reported about unknown model parameters. (Bayarri and Berger, 2004, p. 81)

Would the fact that ‘‘people resist an idea so patently right’’ (Savage, 1954) be fatal to the claim that ‘‘they are Bayesian without knowing it?’’ This is not so sure, experimental investigators could well be right! They feel that the experimental design (incorporating the stopping rule) is prior to the sampling information and that *the information on the design is one part of the evidence*. It is precisely the point of view developed by de Cristofaro (1996, 2004, 2006), who persuasively argued that the correct version of Bayes' formula must integrate the parameter θ, the design d, the initial evidence (prior to designing) e_0, and the statistical information i. Consequently, it must be written in the following form:

$$p(\theta|i, e_0, d) \propto (\theta|e_0, d)p(i|\theta, e_0, d).$$

It becomes evident that the *prior depends on d*. With this formulation, both the likelihood principle and the stopping rule principle are no longer automatic consequences. It is not true that, under the same likelihood, the inference about θ is the same, irrespective of d. Note that the role of the sampling model in the derivation of the Jeffreys prior in Bernoulli sampling for the Binomial and the *Pascal* models was previously discussed by Box and Tiao (1973, pp. 45–46), who stated that the Jeffreys priors are different as the two sampling models are also different. In both cases, the resulting posterior distribution have remarkable frequentist properties (i.e., coverage probabilities of credible intervals).

This result can be extended to general stopping rules (Bunouf, 2006). The basic principle is that the design information, which is ignored in the likelihood function, *can be recovered in the Fisher's information*. Within this framework, we can get a coherent and fully justified Bayesian answer to the issue of sequential analysis, which furthermore satisfy the experimental investigators desideratum (Bunouf and Lecoutre, 2006).

References

Agresti, A., Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics* **61**, 515–523.

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

Albert, J., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Battan, L.J., Neyman, J., Scott, E.L., Smith, J.A. (1969). Whitetop experiment. *Science* **165**, 618.

Baum, M., Houghton, J., Abrams, K.R. (1989). Early stopping rules: clinical perspectives and ethical considerations. *Statistics in Medicine* **13**, 1459–1469.

Bayarri, M.J., Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58–80.

Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 1–17.

Berger, J.O., Bernardo, J.M. (1992). On the development of reference priors (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*. Oxford University Press, Oxford, pp. 35–60.

Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *The American Statistician* **50**, 7–13.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **41**, 113–147.

Bernardo, J.M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences* **4**, 111–121.

Bernardo, J., Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, New York.

Bernoulli, J. (1713). *Ars Conjectandi* (English translation by Bing Sung as Technical report No. 2 of the Department of Statistics of Harvard University, February 12, 1966), Basel, Switzerland.

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

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

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

Brown, L.D., Cai, T., DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science* **16**, 101–133.

Bunouf, P. (2006). *Lois Bayesiennes a priori dans un Plan Binomial Sequentiel*. Unpublished Doctoral Thesis in Mathematics, Université de Rouen, France.

Bunouf, P., Lecoutre, B. (2006). Bayesian priors in sequential binomial design. *Comptes Rendus de L'Academie des Sciences Paris, Série I* **343**, 339–344.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley, Chichester.

Copas, J.B., Loeber, R. (1990). Relative improvement over chance (RIOC) for 2 × 2 tables. *British Journal of Mathematical and Statistical Psychology* **43**, 293–307.

Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

de Cristofaro, R. (1996). L'influence du plan d'echantillonnage dans inférence statistique. *Journal de la Société Statistique de Paris* **137**, 23–34.

de Cristofaro, R. (2004). On the foundations of likelihood principle. *Journal of Statistical Planning and Inference* **126**, 401–411.

de Cristofaro, R. (2006). Foundations of the 'objective Bayesian inference'. In: *First Symposium on Philosophy, History and Methodology of ERROR*. Virginia Tech., Blacksburg, VA.

de Finetti, B. (1972). *Probability, Induction and Statistics: The Art of Guessing*. Wiley, London.

de Finetti, B. (1974). *Theory of Probability* Vol. 1, Wiley, New York.

Dey, D., Rao, C.R. (eds.) (2005). Handbook of Statistics, 25, Bayesian Thinking, Modeling and Computation. Elsevier, North Holland.

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

Dumouchel, W. (1990). Bayesian meta-analysis. In: Berry, D. (Ed.), *Statistical Methodology in Pharmaceutical Science*. Marcel-Dekker, New York, pp. 509–529.

Efron, B. (1998). R.A. Fisher in the 21st century [with discussion]. *Statistical Science* **13**, 95–122.

FDA. (2006). *Guidance for the Use of Bayesian Statistics in Medical Device, Draft Guidance for Industry and FDA Staff*. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Rockville MD.

Fisher, R.A. (1990/1925). *Statistical Methods for Research Workers* (Reprint, 14th ed., 1925, edited by J.H. Bennett). Oxford University Press, Oxford.

Fisher, R.A. (1990/1973). *Statistical Methods and Scientific Inference* (Reprint, 3rd ed., 1973, edited by J.H. Bennett). Oxford University Press, Oxford.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. Wiley, New York.

Fraser, D.A.S., Reid, N., Wong, A., Yi, G.Y. (2003). Direct Bayes for interest parameters. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), *Bayesian Statistics 7*. Oxford University Press, Oxford, pp. 529–534.

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

Gamerman, D. (1997). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall, London.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Good, I.J. (1980). Some history of the hierarchical Bayesian methodology. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), *Bayesian Statistics*. Valencia University Press, Valencia, pp. 489–519.

Guttman, L. (1983). What is not what in statistics? *The Statistician* **26**, 81–107.

Inoue, L.Y.T., Berry, D.A., Parmigiani, G. (2005). Relationship between Bayesian and frequentist sample size determination. *The American Statistician* **59**, 79–87.

Irony, T.Z., Pennello, G.A. (2001). Choosing an appropriate prior for Bayesian medical device trials in the regulatory setting. In: *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*. American Statistical Association, Alexandria, VA.

Iversen, G.R. (2000). Why should we even teach statistics? A Bayesian perspective. In: *Proceedings of the IASE Round Table Conference on Training Researchers in the Use of Statistics*, The Institute of Statistical Mathematics, Tokyo, Japan.

Jaynes, E.T. (2003). In: Bretthorst, G.L. (Ed.), *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Clarendon, Oxford (1st ed.: 1939).

Laplace, P.-S. (1986/1825). *Essai Philosophique sur les Probabilités* (Reprint, 5th ed., 1825). Christian Bourgois, Paris (English translation: *A Philosophical Essay on Probability*, 1952, Dover, New York).

Lecoutre, B. (1996). *Traitement statistique des donnees experimentales: des pratiques traditionnelles aux pratiques bayésiennes* [*Statistical Analysis of Experimental Data: From Traditional to Bayesian Procedures*]. DECISIA, Levallois-Perret, FR (with Windows Bayesian programs by B. Lecoutre and J. Poitevineau, freely available from the web site: http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris).

Lecoutre, B. (2000). From significance tests to fiducial Bayesian inference. In: Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (Eds.), *New ways in statistical methodology: from significance tests to Bayesian inference (2nd ed.)*. Peter Lang, Bern, pp. 123–157.

Lecoutre, B. (2006a). Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science* **4**, 207–232.

Lecoutre, B. (2006b). And if you were a Bayesian without knowing it? In: Mohammad-Djafari, A. (Ed.), *26th Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Conference Proceedings Vol. 872*, Melville, pp. 15–22.

Lecoutre, B., Charron, C. (2006b). Bayesian procedures for prediction analysis of implication hypotheses in 2 × 2 contingency tables. *Journal of Educational and Behavioral Statistics* **25**, 185–201.

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

Lecoutre, B., ElQasyr, K. (2005). Play-the-winner rule in clinical trials: models for adaptative designs and Bayesian methods. In: Janssen, J., Lenca, P. (Eds.), *Applied Stochastic Models and*

Data Analysis Conference 2005 Proceedings, Part X. Health. ENST Bretagne, Brest, France, pp. 1039–1050.

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

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

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

Lee, P. (2004). *Bayesian Statistics: An Introduction*, 3rd ed. Oxford University Press, New York.

Mossman, D., Berger, J. (2001). Intervals for post-test probabilities: a comparison of five methods. *Medical Decision Making* **21**, 498–507.

O'Hagan, A. (1996). *First Bayes* [Teaching Package for Elementary Bayesian Statistics]. Retrieved January 10, 2007, from http://www.tonyohagan.co.uk/1b/.

Pagano, R.R. (1990). *Understanding Statistics in the Behavioral Sciences*, 3rd ed. West, St. Paul, MN.

Rice, W.R. (1988). A new probability model for determining exact P value for 2×2 contingency tables. *Biometrics* **44**, 1–22.

Rosnow, R.L., 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. (2000a). Statistics for researchers. In: Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (Eds.), *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference (2nd ed.)*. Peter Lang, Bern, pp. 1–27.

Rouanet, H. (2000b). Statistical practice revisited. In: Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (Eds.), *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference (2nd ed.)*. Peter Lang, Bern, pp. 29–64.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (2000). *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference*, 2nd ed. Peter Lang, Bern.

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

SAS Institute Inc. (2006). *Preliminary Capabilities for Bayesian Analysis in SAS/STAT® Software*. SAS Institute Inc, Cary, NC.

Savage, L. (1954). *The Foundations of Statistical Inference*. Wiley, New York.

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

Smith, A. (1995). A conversation with Dennis Lindley. *Statistical Science* **10**, 305–319.

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

Sweeting, T.J. (2005). On the implementation of local probability matching priors for interest parameters. *Biometrika* **92**, 47–57.

Tan, S.B., Chung, Y.F.A., Tai, B.C., Cheung, Y.B., Machin, D. (2003). Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. *Controlled Clinical Trials* **24**, 110–121.

Toecher, K.D. (1950). Extension of the Neyman–Pearson theory of tests to discontinuous variables. *Biometrika* **37**, 130–144.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles [with discussion]. *Journal of the Royal Statistical Society B* **58**, 3–57.

Winkler, R.L. (1974). Statistical analysis: theory versus practice. In: Stael Von Holstein, C.-A.S. (Ed.), *The Concept of Probability in Psychological Experiments*. D. Reidel, Dordrecht, pp. 127–140.

Zaykin, D.V., Meng, Z., Ghosh, S.K. (2004). Interval estimation of genetic susceptibility for retrospective case–control studies. *BMC Genetics* **5**(9), 1–11.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* **64**, 131–146.