And if you were a Bayesian without knowing it?

Bruno Lecoutre

ERIS and UPRESA 6085, Laboratoire de Mathématiques Raphaël Salem
C.N.R.S. et Université de Rouen
Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray, France
e-mail: bruno.lecoutre@univ-rouen.fr
`http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris`

Abstract. The literature is full of Bayesian interpretations of frequentist p-values and confidence levels. All the attempts to rectify these interpretations have been a loosing battle. In fact such interpretations suggest that most users are likely to be Bayesian "without knowing it" and really want to make a different kind of inference.

Keywords: Bayesian inference, Frequentist inference, Confidence intervals, Inverse probabilities
PACS: 01.70.+w, 02.50.Cw, 02.50.Tt

INTRODUCTION

Many statistical users misinterpret the p-values of significance tests as inverse probabilities: $1-p$ is "the probability that the alternative hypothesis is true". As is the case with significance tests, the frequentist interpretation of a 95% confidence interval (CI) involves a long run repetition of the same experiment: in the long run 95% of CI 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 the "correct" interpretation of CIs does not make sense for most users. Paradoxically it is the interpretation in (Bayesian) terms of "a *fixed* interval having a 95% chance of including the true value of interest" which is their appealing feature.

All the attempts to rectify these misinterpretations have been a loosing battle. In particular, virtually all users interpret frequentist confidence intervals in a Bayesian fashion. Consequently we automatically ask themselves: "**and if you were a Bayesian without knowing it?**"

FREQUENTIST AND BAYESIAN INFERENCE

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, that seems to make probability an observable 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.

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. Moreover, 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.

The frequentist approach 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, and confidence intervals. In the Bayesian inference parameters can also be probabilized. This results in distributions of probabilities that express our uncertainty: (1) before observations (they do not depend on data): *prior* probabilities; (2) after observations (conditional on data): *posterior* (or *revised*) probabilities; (3) about future data: *predictive* probabilities.

As a simple illustration let us consider the following situation. A finite population of size 20 with a dichotomous variable success/failure and a proportion φ of success. Hence the unknown parameter φ. A sample of size five has been observed. Hence the *known data*: 0 0 0 1 0 ($f = 1/5$). 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.

Two different approaches to statistical inference

The frequentist approach: from unknown to known. In the frequentist framework, we have no probabilities and consequently no possible inference. So 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. These sampling probabilities serve to define a significance test. Given the data in hand , 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: 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, that is irrelevant for my purpose.

However, this conclusion is based on the probability of the samples *that have not been*

observed, what Jeffreys (1998/1939, Section 7.2) ironically expressed in the following terms: "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 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% CI: $[0.05, 0.60]$. 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* !

The Bayesian approach: from known to unknown. 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. We consider, not the frequentist probabilities of imaginary samples, but the frequentist probabilities of *the observed data* ($f = 1/5$) for *all possible values* of the parameter φ. This is the *likelihood function* that is denoted by $\ell(\varphi|data)$.

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:

$$Pr(\varphi \text{ and } f = 1/5) = Pr(f = 1/5 \,|\, \varphi) \times Pr(\varphi) = \ell(\varphi|data) \times Pr(\varphi).$$

The sum of the joint probabilities gives the marginal predictive probability of the data, before observation, which is very intuitive since the predictive probability is a weighted average of the likelihood function, the weights being the prior probabilities:

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

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

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

The frequentist approach involves considerable difficulties. We can conclude with Berry (1997) that "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*. Frequent questions asked by statistical users show us that this use is counterintuitive: "why one considers the probability of samples outcomes that are more extreme than the one observed?"; "why must one calculate the probability of samples that have not been observed?"; etc.

Such difficulties are not encountered with the Bayesian inference: the posterior distribution, being conditional on data, only involves the sampling probability of the data *in hand*, via the likelihood function $\ell(\varphi|data)$ that writes the sampling distribution in the *natural order*: "from unknown to known".

EXPERIMENTAL RESEARCH AND STATISTICAL INFERENCE

We are facing a paradoxical situation. On the one hand, Null Hypothesis Significance Testing (NHST) has been long required in most scientific publications as an unavoidable norm and it often appears as a label of scientificness. But on the other hand, it leads to innumerable misuses (for a review, see e.g. Rouanet *et al.*, 2001; Lecoutre *et al.*, 2001). Furthermore, its use has been explicitly denounced by the most eminent and most experienced scientists, both on theoretical and methodological grounds

Today is a crucial time because we are in the process of defining new publication norms for experimental research. While users' uneasiness is ever growing, changes in reporting experimental results, especially in presenting and interpreting effect sizes, are more and more enforced within editorial policies in all fields.

Users' dissatisfaction with current practices. Even professional applied statisticians from pharmaceutical companies are not immune to misinterpretations of NHST, especially if the test is nonsignificant (Lecoutre *et al.*, 2003). It is hard to interpret this finding as an individual's lack of mastery. Actually, this reveals that NHST does not address questions that are of primary interest for the scientific research. A statistically significant test provides no information about the departure from the null hypothesis; with a large sample a descriptively small departure may be significant. A nonsignificant test is not evidence favouring the null hypothesis: for an insufficiently sensitive experiment descriptively large departure from the null hypothesis may be nonsignificant.

In fact, in order to interpret their data in a reasonable way, users must resort to a more or less "naive" mixture of NHST results and other information. But this is not an easy task! Actually, many researchers explicitly state that they are dissatisfied with current practices and appear to have a real consciousness of the stranglehold of NHST. They use significance tests only because they know no other alternative, but they express the need for inferential methods that would be better suited for answering their specific questions.

A set of recipes and rituals. There are currently many attempts to remedy the inadequacy of NHST. In particular, the necessity of reporting effect size estimates and their confidence intervals is stressed. The role of the planning of experiments (how many subjects to use) is also emphasized and power computations are recommended. In fact, these attempts are both partially technically redundant and conceptually incoherent. Just as NSHT, they should result in a set of *recipes and rituals*, without supplying *a real statistical thinking*. In particular, one can be afraid that statistical users continue to focus on the statistical significance of the result (only wondering whether the CI includes the null hypothesis value) rather than on the full implications of confidence intervals.

New difficulties with confidence intervals. CIs could quickly become a compulsory norm in experimental publications. 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 *noninformative* 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). Then the debates can be expressed on these terms: "whether the probabilities should only refer to data and be based on frequency or whether they should also apply to parameters and be regarded as measures of beliefs."

The ambivalence of statistical instructors. For many reasons due to their frequentist conception, CIs can hardly be viewed as the ultimate method. It is undoubtedly their intuitive (Bayesian) interpretation in terms of "a fixed interval having a 95% chance of including the true value of interest" which is their appealing feature. Ironically these heretic interpretations are encouraged by the ambivalence of most statistical instructors who tolerate and even use them.

So, in a popular statistical textbook (Pagano, 1990, page 288), that claims the goal of "*understanding statistics*", a 95% CI is described as "*an interval such that the probability is* 0.95 *that the interval contains the population value*". Other authors claim that the correct frequentist interpretation they advocate can be expressed as: "we can be 95% confident that the population mean is between 114.06 and 119.94" (Kirk, 1982, page 43). It is hard to imagine that the reader can understand that "confident" refers here to a frequentist view of probability!

THE BAYESIAN ALTERNATIVE

It is not acceptable that that future statistical inference methods users will continue using non appropriate procedures *because they know no other alternative*. Since most people use "inverse probability" statements to interpret NHST and CIs, the Bayesian approachisis, at least implicitly, involved in the use of frequentist methods. Which is simply required by is a very natural shift of emphasis about the concepts of Bayesian inference, showing that they can be used consistently and appropriately in statistical analysis (Lecoutre, 2006).

With the Bayesian approach, 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.

A better understanding of frequentist procedures. As a simple illustration of how the Bayesian procedures combine descriptive statistics and significance tests, let us consider the basic situation of the inference about the difference δ between two normal means. 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 noninformative 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, 2006).

Conceptual links result from this *technical* link. 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 CI 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). This interval is usually termed a Bayesian *credible interval*, which explicitly accounts for the difference in interpretation.

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 CIs in the language of probabilities about unknown effects is quite natural for 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".

The use of noninformative priors has a privileged status in order to gain "public use" statements. Of course, when "good prior information is available" other Bayesian techniques also have an important role to play in experimental investigations. They are ideally suited for *combining information* from the data in hand and from other studies.

Bayesian procedures are no more arbitrary than frequentist ones. Many potential users continue to think that Bayesian methods are too subjective to be scientifically acceptable. However, frequentist methods are full of more or less *ad hoc* conventions. Thus the p-value is 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. For instance, let us consider 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*.

Obviously, in this case the choice of a noninformative prior distribution cannot avoid conventions, but it is an exact counterpart of the arbitrariness involved within the frequentist approach. For Binomial sampling, different priors have been proposed for an objective Bayesian analysis. In fact, it exists two extreme noninformative priors that are respectively the more unfavourable and the more favourable 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. 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.

Then the usual criticism of frequentists towards the divergence of Bayesians with respect to the choice of a noninformative 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 Jeffreys prior credible interval has remarkable frequentist properties. Its coverage probability is very close to the nominal level, even for small-size samples. It is undoubtedly an objective procedure that can be favourably compared to most frequentist intervals.

Similar results are obtained for negative-Binomial (or Pascal) sampling. 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 to privilege the intermediate Beta distribution of parameters

0 and $1/2$, which is precisely the Jeffreys prior. This result concerns a very important issue related to the "likelihood principle" that I shall address in more detail further on.

The predictive probabilities: A very appealing tool. A major strength of the Bayesian paradigm is the ease to make predictions about future observations. The predictive idea is central in experimental investigations as the essence of science is replication. 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.

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. This desideratum is currently considered 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, page 81).

Would the fact that "people resist an idea so patently right" 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 (2004), 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:

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

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.

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, pages 45–46), who stated that the Jeffreys priors are different as the two sampling models are also different. 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 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).

CONCLUSION

I suggest that an "objective Bayes theory" is by no means a speculative viewpoint but on the contrary is perfectly feasible (Lecoutre *et al.*, 2001; Berger, 2004). It is better suited to the needs of users than frequentist approach and provide scientists with relevant answers to essential questions raised by experimental data analysis. 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 fact the times we are living in at the moment appear to be crucial. One of the decisive factors could be the recent "draft guidance document" of the US Fud 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.

REFERENCES

1. M. J. Bayarri, and J. O. Berger, "The interplay of Bayesian and frequentist analysis," *Statistical Science* **19**, 58–80 (2004).
2. J. Berger, "The case for objective Bayesian analysis," *Bayesian Analysis* **1**, 1–17 (2004).
3. D. A. Berry, "Teaching elementary Bayesian statistics with real applications in science," *The American Statistician* **51**, 241–246 (1997).
4. G. E. Box, and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Wiley, New York, 1973.
5. P. Bunouf, *Lois Bayésiennes a Priori dans un Plan Binomial Séquentiel*, unpublished doctoral thesis in mathematics, Université de Rouen, 2006.
6. P. Bunouf, and B. Lecoutre. "Bayesian priors in sequential binomial design," *Comptes Rendus de L'Académie des Sciences Paris, Série I* **343**, 339–344.
7. R. de Cristofaro, "On the foundations of likelihood principle," *Journal of Statistical Planning and Inference* **126** 401–411 (2004).
8. FDA, *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, May 22 2006.
9. E. T. Jaynes, *Probability Theory: The Logic of Science* (Edited by G. L. Bretthorst), Cambridge University Press, Cambridge, England, 2003.
10. H. Jeffreys, *Theory of Probability*, Clarendon, Oxford (1st edn: 1939), 1998 (3rd edn).
11. R E. Kirk, *Experimental Design. Procedures for the Behavioral Sciences*, Brooks /Cole, Pacific Grove, CA, 1982.
12. B. Lecoutre, "Training students and researchers in Bayesian methods for experimental data analysis," *Journal of Data Science* **4**, 207–232 (2006).
13. B. Lecoutre, M.-P. Lecoutre, and J. Poitevineau, "Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable?" *International Statistical Review* **69**, 399–418 (2001).
14. M.-P. Lecoutre, J. Poitevineau, and B. Lecoutre, "Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests," *International Journal of Psychology* **38**, 37–45 (2003).
15. R. R. Pagano, *Understanding statistics in the behavioral sciences*, West, St. Paul, MN, 1990 (3rd edn).
16. H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre, and B. Le Roux, *New ways in statistical methodology: From significance tests to Bayesian inference*, Peter Lang, Bern, SW, 2000 (2nd edn).