Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, Spain.
Tel. +34.96.364.3560 (direct), +34.96.386.4362 (office).
Fax +34.96.364.3560 (direct), +34.96.386.4735 (office).
Internet: jose.m.bernardo@uv.es, Web: http://www.uv.es/˜bernardo/

Typesetted on October 3, 2002
To appear at *Internat. Statist. Rev.*

A Bayesian Approach to some Cryptic Issues on the Nature of Statistical Inference

JOSÉ M. BERNARDO
Universitat de València, Spain
<jose.m.bernardo@uv.es>

SUMMARY

This note summarizes the basic argument for a Bayesian approach to statistical inference and, from that perspective, provides a possible answer to each of the fourteen 'cryptic' issues on the nature of statistical inference formulated by Sir David Cox within his 1997 Bernoulli Lecture at the University of Groningen.

Keywords: ADMISSIBILITY; LOSS FUNCTION; OBJECTIVE BAYESIAN PROCEDURES; REFERENCE ANALYSIS.

On the occasion of his 1997 Bernoulli Lecture at Groningen University (The Netherlands), Sir David Cox exercised his usual wit to propose consideration of 14 cryptic issues on the nature of statistical inference. A group of scholars (Kardaun *et al.*, 2002) formulated some sort of *commmunis opino* on these matters from a traditional, mostly frequentist, viewpoint. In a somewhat telegraphic style, a possible alternative approach to those issues is described below from very different viewpoint. In Section 1, some required foundational points are discussed; Section 2 contain a possible Bayesian approach to the 14 issues discussed.

1. FOUNDATIONS

(i) Bayesian decision theory . Established on a solid mathematical basis, Bayesian decision theory provides a privileged platform to coherently discuss basic issues on statistical inference. Indeed, even from a strictly frequentist perspective, most purely inferential problems are best analyzed as decision problems under uncertainty. Thus, for data $\boldsymbol{z} \in \mathcal{Z}$ whose probabilistic behaviour is assumed to be described by some probability model $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, any statistical procedure may be identified with some (possibly complicated) function $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z}) \in \mathcal{T}$ (where $\mathcal{T}$ may well be a function space). Obvious examples include a point estimator, a confidence region, a test procedure or a posterior distribution. For each particular procedure, it should be possible to define a *loss function* $L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}$ which somehow measures the 'error'

José M. Bernardo is Professor of Statistics at the University of Valencia, Spain. Research partially funded with grants GV01-7 of the Generalitat Valenciana and BNF2001-2889 of the DGICYT, Madrid, Spain

committed by the procedure as a function of $\boldsymbol{\theta}$. Usual loss functions include the quadratic loss $L\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$ associated to a point estimator $\tilde{\boldsymbol{\theta}}$ and the logarithmic loss $L\{\pi_\theta(.\,|\,\boldsymbol{z}), \boldsymbol{\theta}\} = -\log[\pi(\boldsymbol{\theta}\,|\,\boldsymbol{z})]$ associated to a posterior density $\pi_\theta(.\,|\,\boldsymbol{z})$.

Conditional on observed data $\boldsymbol{z}$, the Bayes procedure $\boldsymbol{t}^b(\boldsymbol{z})$ which corresponds to a proper prior $\pi(\boldsymbol{\theta})$ is that minimizing the corresponding posterior loss

$$\boldsymbol{t}^b(\boldsymbol{z}) = \arg \inf_{\boldsymbol{t} \in \mathcal{T}} \int_\Theta L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\} \pi(\boldsymbol{\theta}\,|\,\boldsymbol{z})\, d\boldsymbol{\theta}, \quad \pi(\boldsymbol{\theta}\,|\,\boldsymbol{z}) \propto p(\boldsymbol{z}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

A procedure $\boldsymbol{t}^*(\boldsymbol{z})$ is a *generalized Bayes procedure* if there exists a sequence $\{\pi_n(\theta)\}$ of proper priors yielding a sequence of Bayes procedures $\{\boldsymbol{t}_n^b(\boldsymbol{z})\}$ such that $\boldsymbol{t}^*(\boldsymbol{z}) = \lim_{n\to\infty} \boldsymbol{t}_n^b(\boldsymbol{z})$.

(ii) Admissibility Conditional on $\boldsymbol{\theta}$ and considered as a function of the data $\boldsymbol{z}$, the loss function $L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}$ is a random quantity, whose expectation (under repeated sampling),

$$R_t(\boldsymbol{\theta}\,|\,L) = \mathrm{E}_{z\,|\,\theta}[L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}] = \int_{\mathcal{Z}} L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}\, p(\boldsymbol{z}\,|\,\boldsymbol{\theta})\, d\boldsymbol{z},$$

provides a a description of the *average risk* involved in using the procedure $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ as a function of the unknown parameter vector $\boldsymbol{\theta}$. A relatively small average risk $R_t(\boldsymbol{\theta}\,|\,L)$ with respect to reasonable loss functions L is certainly a *necessary* condition for the procedure $\boldsymbol{t}$ to be sensible, but it is hardly sufficient: the procedure may well have an unacceptable behaviour with specific data $\boldsymbol{z}$ and yet produce an small average risk, either because those data are not very likely, or because errors are somehow averaged out.

When comparing the risks associated to two alternative procedures designed to perform the same task, it may well happen that (with respect to a particular loss function L) a procedure $t_1(\boldsymbol{z})$ is *uniformly* better than another procedure $t_2(\boldsymbol{z})$ in the sense that $\forall \boldsymbol{\theta} \in \Theta$, $R_{t_1}(\boldsymbol{\theta}\,|\,L) < R_{t_2}(\boldsymbol{\theta}\,|\,L)$; it is then said that $t_2(\boldsymbol{z})$ is *dominated* by $t_1(\boldsymbol{z})$, and $t_2(\boldsymbol{z})$ is declared to be *inadmissible* with respect to that loss function. A crucial, too often ignored result (Savage, 1954; Berger, 1985, Ch. 8, and references therein) says however that, under suitable regularity conditions, a *necessary and sufficient* condition for a procedure to be admissible is to be a generalized Bayes procedure. It follows that, even from a purely frequentist viewpoint, one should strive for (generalized) Bayes procedures.

(iii) Objective Bayesian Procedures As Kardaun *et al.* (2002) (and many people before them) point out, one role of statistical theory is to provide a broadly acceptable framework of concepts and methods which may be used to provide a ‘professional’ answer. If it may reasonably be assumed that the probability model $\{p(\boldsymbol{z}\,|\,\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ encapsulates *all* objective available information on the probabilistic structure of the data, then such a professional answer should not depend on a subjectively assessed prior $\pi(\boldsymbol{\theta})$. Note, that structural assumptions on the data behaviour (such as partial exchangeability) are easily accommodated within this framework; one would then have some form of *hierarchical model*, $\{p(\boldsymbol{z}\,|\,\boldsymbol{\phi}), \pi(\boldsymbol{\phi}\,|\,\boldsymbol{\theta})\}$, where $\boldsymbol{\theta}$ would be a hyperparameter vector, $p(\boldsymbol{z}\,|\,\boldsymbol{\theta}) = \int_\Phi p(\boldsymbol{z}\,|\,\boldsymbol{\phi})\pi(\boldsymbol{\phi}\,|\,\boldsymbol{\theta})\, d\boldsymbol{\phi}$ would be the corresponding ‘integrated’ model, and a prior $\pi(\boldsymbol{\theta})$ would be required for the original hyperparameters.

An objective Bayesian procedure to draw inferences about some quantity of interest $\phi = \phi(\boldsymbol{\theta})$, requires an objective ‘non-informative’ prior (‘objective’ in the precise sense that it exclusively depends on the the assumed model $\{p(\boldsymbol{z}\,|\,\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ and the quantity of interest), which mathematically describes lack on relevant information about the quantity of interest ϕ. The statistical literature contains a number of requirements which may be regarded as necessary properties of any algorithm proposed to derive these ‘baseline’ priors; those requirements include general applicability, invariance under reparametrization, consistent marginalization, and appropriate coverage properties. The *reference analysis* algorithm, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1992), provides a general method to derive objective priors which apparently satisfies all these desiderata, and which is shown to contain many previous results (*e.g.*, maximum entropy and univariate Jeffreys' rule) as particular cases.

Reference priors are defined as a limiting form of proper priors (obtained by maximization of an information measure), and are shown to yield generalized Bayes procedures. Thus, reference analysis may be used to obtain *objective* Bayesian solutions which show both appropriate conditional properties (for they condition on the actual, observed data) and an appealing behaviour under repeated sampling (for they are typically admissible).

2. THE CRYPTIC ISSUES

A possible Bayesian answer is now provided to each of the fourteen issues under discussion. Unless otherwise indicated, the statements made are valid whatever the procedure used to specify the prior: objective (model-based), or subjectively assessed.

1. How is overconditioning to be avoided? Both overconditioning and overfitting are aspects of inappropriate model choice. Model choice is best described as a decision problem where the action space is the class of models $\{M_i \in \mathcal{M}\}$ which one is prepared to consider, and its solution requires specifying a loss function which measures, as a function of the quantity of interest $\boldsymbol{\phi}$, the consequences $L(M_i, \boldsymbol{\phi})$ of using a particular model M_i within the specific context one has in mind.

For instance, if given a random sample $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ one is interested in prediction of a future observation x, an appropriate loss function might be written in terms of the logarithmic scoring rule, so that $L\{M_i, x\} = -\log\{p_i(x \mid z)\}$, and the best available model would be that which minimizes within $\mathcal{M}$ the corresponding (posterior) expected loss,

$$\overline{L}(M_i \mid \boldsymbol{z}) = \int_{\mathcal{X}} L\{M_i, x\}\, p(x \mid \boldsymbol{z})\, dx = -\int_{\mathcal{X}} p(x \mid \boldsymbol{z}) \log\{p_i(x \mid z)\, dx,$$

a predictive cross-entropy. Since the true model, and hence the true predictive density $p(x \mid \boldsymbol{z})$, are not known, some form of approximation is necessary; direct Monte Carlo approximation to the integral above leads to

$$\overline{L}(M_i \mid \boldsymbol{z}) \approx -\frac{1}{n} \sum_{j=1}^{n} \log\{p_i(x_j \mid z_j), \quad \boldsymbol{z}_j = \boldsymbol{z} - \{x_j\},$$

closely related to cross-validation techniques; (for details, see Bernardo and Smith, 1994, Ch. 6).

2. How convincing is the likelihood principle? An immediate consequence of Bayes theorem is that, *conditional* to a given prior $\pi(\boldsymbol{\theta})$, the posterior distributions obtained from proportional likelihoods are identical. In this limited sense, the likelihood 'principle' is an obvious consequence of probability theory, and any statistical procedure which violates this should not be trusted. That said, there are many statistical problems which require consideration of the sample space and will, therefore, typically yield different answers for different models, even if those happen to yield proportional likelihood functions. Design of experiments, a decision problem where the best experiment within a given class must be chosen, or prediction problems, where the predictive posterior distribution of some future observables must be found, are rather obvious examples. The likelihood principle should certainly not be taken to imply that the sample space is irrelevant.

Objective Bayesian inference in another example where the sample space matters. The reference prior is defined as that which maximizes the missing information about the quantity of interest which could be provided by the experiment under consideration; thus, different probability models, even those with proportional likelihood functions, will generally yield different reference priors. For instance, the reference prior which corresponds to binomial sampling is $\pi^*(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, but the reference prior which corresponds to inverse binomial sampling is $\pi^*(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$, a difference which reflects the fact that in the second case one is implicitly assuming that one success will eventually be observed; for details, see Bernardo and Smith, 1994, Ch. 5.

3. What is the role of probability in formulating models where hypothetical repetition is hard to envisage? Probability is a measure on degree of rational belief conditional on any available information. This concept of probability does not require symmetries or hypothetical repetitions, although it obviously will take these into account as relevant information when available.

We certainly agree with the statement by Kardaun *et al.* (2002) that algorithms 'should work well with simulated data'. Indeed, when teaching Bayesian methods, it is important to display the results from simulated data for the student to see that, as one would certainly expect, the posterior density of any parameter concentrates around its true value, or the predictive distribution of a future observation approaches the true model. That said, the frequentist interpretation of probability is simply too narrow for many important applications of statistics. Probability may *always* be interpreted in its epistemological sense in ordinary language: as a conditional measure or rational belief.

4. Should nonparametric and semiparametric formulations be forced into a likelihood-based framework? Nonparametrics is something of a misnomer. When a model is assumed to be of the form $\{p(\boldsymbol{z}\,|\,\boldsymbol{\theta}), \boldsymbol{\theta}\in\Theta\}$ nothing is said about the nature of the parameter space Θ, which may well label, say, the class of all absolutely continuous densities of $\boldsymbol{z}\in\mathcal{Z}$. It is just a convention to call 'parametric' those problems where $\Theta\subset\Re^k$, so that $\boldsymbol{\theta}$ is a vector of finite dimension k. Whether or not it is better to use a 'nonparametric' (infinitely dimensional) formulation, certainly more general but requiring a prior distribution defined on a function space, than it is to work with a model labeled by a parameter with finite dimension is just another example of a problem of model choice, to which the comments in (i) above are directly relevant.

5. Is is fruitful to treat inference and decision analysis somewhat separately? At a foundational level certainly it is *not*: decision analysis provides the coherent framework which guarantees that no inconsistencies and/or obviously wrong answers (a *negative* unbiased estimate of a probability, or a 95% confidence region for a real-valued quantity which happens to be the *entire* real line, say), will be derived. For an interesting collection of counterexamples to conventional frequentist methods see Jaynes (1976) and references therein.

That said, it is important to formalize those situations where 'pure' inference is of interest, as opposed to specific (context dependent) decision problems. This is easily done however within the context of decision analysis: for instance, pure, abstract inference on the value of $\boldsymbol{\theta}$ may be described a decision problem on the best way to describe the posterior uncertainty on the value of $\boldsymbol{\theta}$, where the value of the consequences are described with an information measure (Bernardo, 1979a). The simple decision-theoretical formulation of most text-book statistical procedures is well known: point estimation is best seen as a decision problem where the action space is the parameter space; testing a null hypothesis $H_0\equiv\{\boldsymbol{\theta}\in\Theta_0\}$ is best seen as a decision problem on whether or not to work as if $\boldsymbol{\theta} \in \Theta_0$. Practical application of these ideas require however the identification of appropriate loss functions; for some new ideas in this area, see Bernardo and Rueda (2002) and Bernardo and Juárez (2003).

6. How possible and fruitful is to treat qualitatively uncertainty not derived from statistical variability? In statistical consulting one is routinely forced to consider uncertainties which are *not* derived from statistical variability, and hence 'professional' statistical answers must be able to deal with them. This is naturally done within a Bayesian framework, and simply *cannot* be done within a frequentist framework. Other approaches to quantify uncertainty (belief functions, discord,...) have so far failed to provide a *consistent* mathematical framework which could be used instead of probability theory to measure and to operate with uncertainty.

7. Are all sensible probabilities ultimately frequency based? Although one could certainly use a different word than probability for a rational degree of belief (say 'credence' or 'verisimilitude') this is not really needed: mathematically, probability is a well-defined concept, a measure function endowed with certain properties, and the foundations of decision theory prove that degrees of belief *must* have this mathematical structure; hence they *are* probabilities in the mathematical sense of the word.

That said, the important frequentist interpretation of probability models based on the concept of *exchangeability* (de Finetti, 1937; Hewitt and Savage, 1955; Bernardo and Smith, 1994, Ch. 4) is often neglected: *all random samples are necessarily exchangeable* and, by virtue of the probability theory-based general *representation theorem*, the parameter identifying the model which describes its probabilistic behaviour is *defined* in terms of the long-term behaviour of some function of the observations. Thus, a set of exchangeable Bernoulli observations is *necessarily* a random sample of dichotomous observations with common parameter θ, *defined* as the long-term limit of the relative frequency of successes.

The representation theorem further establishes the *existence* of a prior $\pi(\boldsymbol{\theta})$ for the parameter, so that (whenever one has exchangeable observations, and—to insist—all random samples are exchangeable) the frequently heard sentence 'there is no prior distribution' is simply *incompatible* with probability theory.

8. Was R. A. Fisher right to deride axiomatic formulations in statistics? If he did, he was entirely wrong. Ever since classical Greece, mathematicians have strived to provide axiomatic foundations on their subject as a guarantee of self-consistency. By the early 20th century this process had been completed in all mathematical branches (including probability theory) except mathematical statistics. No wonder that contradictions arose in conventional statistics, and no surprise at the often derogatory attitude of mathematicians to mathematical statistics, too often presented as an 'art' where contradictions could be acknowledged and were to be decided by the wit of the 'artist' statistician.

To argue that axiomatics 'ought not to be taken seriously in a subject with real applications in view' (is geometry void of real applications?), just because new concepts might be necessary is to ignore how science progresses. A paradigm is obviously only valid until it cannot explain new facts; then a new, self-consistent paradigm must be found (Kuhn, 1962). The frequentist paradigm is simply *not* sufficient for present day applications of statistics; at least today, the Bayesian paradigm is.

It may well be that alternative axiomatic basis for mathematical statistics are possible beyond that provided by decision theory (which leads to a Bayesian approach), although none has been presented so far. But, whether or not alternatives appear, statistical inference should not be deprived of the mathematical firmware provided by sound foundations; or would anyone trust an architect trying to build a beautiful house on shaky foundations?

9. How can randomization be accommodated within statistical theory? Randomization is not necessary for a single decision maker: if a_1 and a_2 (which may well be two alternative designs) have expected utilities $\overline{U}(a_1 \,|\, \boldsymbol{z})$ and $\overline{U}(a_2 \,|\, \boldsymbol{z})$, the randomized action which takes a_1 with probability γ and a_2 with probability $1-\gamma$, $(0 \leq \gamma \leq 1)$ has an expected utility given by $\gamma\overline{U}(a_1 \,|\, \boldsymbol{z}) + (1-\gamma)\overline{U}(a_2 \,|\, \boldsymbol{z})$, so that randomization for a single decision maker could apparently only be optimal if $\overline{U}(a_1 \,|\, \boldsymbol{z}) = \overline{U}(a_2 \,|\, \boldsymbol{z})$ and then only as good as a non-randomized action. The situation is however very different if more than one decision maker is involved. As suggested by Stone (1969), randomization becomes optimal if the decision maker takes into account that he/she has to convince other people, not just him/her self. For details, see Berry and Kadane (1997).

10. Is the formulation of personalistic probability by de Finetti and Savage the wrong way round? Kardaun *et al.* (2002) (and, again, many before them) suggest that there are no compelling reasons for epistemic probabilities to behave as mathematical probabilities. Yet, it is difficult to imagine something more compelling than a mathematical proof; it is *not* simply that it makes intuitive sense to use probabilities: the fact is that behaviour of epistemic probabilities as mathematical probabilities follows from rather intuitive axioms on coherent behaviour. Kardaun *et al.* (2002) seem to be happy with the rather obvious inconsistencies of the conventional paradigm when they say 'the incoherent behaviour thus displayed will hopefully lead to a higher degree of verisimilitude than the coherent behaviour...'(!); one is lead to wonder how would they react when approached by a car salesman who admits that the car he suggests gets bad mileage both in town conditions and in road conditions, only to claim that it gets a good mileage overall.

11. How useful is a personalistic theory as a base for public discussion? If by personalistic it is meant subjective, not much (although it will have the merit of making explicit peoples' assumptions, which is more than one often gets from public discussion). However, if by personalistic one merely means epistemic probability, *i.e.*, probabilities interpreted as rational degrees of belief conditional only to whatever 'objective' (often meaning intersubjective) assumptions one is prepared to make, then this is precisely the base for public discussion. And this is precisely the type of result that objective Bayesian methods provide.

Indeed, in any given experimental situation, the scientist typically wants to make an inferential claim, say the result $t(\boldsymbol{z})$ of a statistical procedure, conditional on any assumptions made, and given the *observed* data $\boldsymbol{z}$. What might have happened had other data $\boldsymbol{z} \in \mathcal{Z}$ been obtained might relevant to *calibrate* the procedure (see Section 1 above); however, what is actually required to describe the inferential content of the experimental results is a measure on the degree of rational belief on the conclusion advanced, not a measure of the behaviour of the procedure under repeated sampling.

As Kardaun *et al.* (2002) write, 'statisticians are hired... to make scientifically sound statistical inferences in the light of data and in spite of some unavoidable uncertainty'. However, this is precisely what objective Bayesian methods do, but frequentist computations do not.

12. In a Bayesian formulation should priors be constructed retrospectively? From a subjectivist viewpoint construction may proceed either way, provided all conditioning operations are properly done: one may directly assess the posterior distribution! However, subjective probability assessment is a very hard task, and any appropriate mechanism such as Bayes theorem or extending the conversation to include other variables, should be used to help in completing the task.

Note, however, that if one is directly interested in objective Bayesian methods the question does not arise. Given data $\boldsymbol{z} \in \mathcal{Z}$ and model $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, the prior $\pi^*_\phi(\boldsymbol{\theta})$ required to obtain a reference posterior distribution $\pi^*(\phi \mid \boldsymbol{z})$ for a particular quantity of interest $\phi = \phi(\theta)$, that which maximizes the missing information about the value of ϕ, is a well-defined mathematical function of the probability model $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ and the quantity of interest $\phi(\theta)$.

13. Is the only justification of much current Bayesian work using rather flat priors the generation of (approximate) confidence limits? or do the various forms of reference priors have some other viable justification? In their discussion of this issue, Kardaun *et al.* (2002) have unfortunately chosen to ignore over 30 years or research: what they write could have been written as part of one of the many discussions on this topic published in the 60's and 70's. We do not have space here to describe the basics of Bayesian objective methods, let alone to document the huge relevant literature. For a textbook level description of some objective Bayesian methods see Bernardo and Smith (1994, Ch. 5); for critical overviews of the topic, see Kass and Wasserman (1996), Bernardo (1997), references therein and ensuing discussion.

Reference analysis has been mentioned in Section 1 as an advanced procedure to derive objective priors (which, except for location parameters, are certainly not 'flat'). Just to give a hint of the ideas behind, the basic definition of reference priors in the one-parameter case is quoted below.

The amount of information $I^\theta\{\mathcal{Z}, \pi_\theta(.)\}$ which an experiment yielding data $\boldsymbol{z} \in \mathcal{Z}$ may be expected to provide about θ, a function of the prior $\pi_\theta(.)$, is (Shannon, 1948)

$$I^\theta\{\mathcal{Z}, \pi_\theta(.)\} = \int_{\mathcal{Z}} p(\boldsymbol{z}) \int_\Theta \pi(\theta \mid \boldsymbol{z}) \log \frac{\pi(\theta \mid \boldsymbol{z})}{\pi(\theta)}\, d\theta d\boldsymbol{z}.$$

If this experiment were continuously replicated, the true value of θ would eventually be learnt. Thus, the amount of information to be expected from k replicates of the original experiment, $I^\theta\{\mathcal{Z}^k, \pi_\theta(.)\}$, will converge (as $k \to \infty$) to the missing information about θ associated to the prior $\pi_\theta(.)$. Intuitively, the reference prior $\pi^*_\theta(.)$ is that which maximizes the *missing information* about θ within the class of priors compatible with accepted assumptions.

Formally, if $\mathcal{P}$ is the class of accepted priors in the problem considered (which may well be the class of all priors), the reference posterior $\pi^*(\theta \mid \boldsymbol{z})$ is defined by

$$\pi^*(\theta \mid \boldsymbol{z}) = \lim \pi_k(\theta \mid \boldsymbol{z}),$$

where the limit is taken in the (information) sense that

$$\lim_{k\to\infty} \int_\Theta \pi_k(\theta \mid \boldsymbol{z}) \log \frac{\pi_k(\theta \mid \boldsymbol{z})}{\pi(\theta \mid \boldsymbol{z})}\, d\theta = 0,$$

and where $\pi_k(\theta \mid \boldsymbol{z}) \propto p(\boldsymbol{z} \mid \theta)\, \pi_k(\theta)$ is the posterior which corresponds to the prior

$$\pi_k(\theta) = \arg \sup_{\pi(\theta)\in\mathcal{P}} I^\theta\{\mathcal{Z}^k, \pi(\theta)\}.$$

maximizing in $\mathcal{P}$ the amount of information to be expected from k replicates of the original experiment. Finally, a reference 'prior' is any positive function $\pi^*(\theta)$ such that

$$\pi^*(\theta \mid \boldsymbol{z}) \propto p(\boldsymbol{z} \mid \theta)\, \pi^*(\theta),$$

so that the reference posterior may be simply obtained by formal use of $\pi^*(\theta)$ as a (typically improper) prior.

It may be proved that if the parameter space Θ is finite this leads to maximum entropy and if Θ is non-countable, $p(\boldsymbol{z} \mid \theta)$ regular and $\mathcal{P}$ is the class of all strictly positive priors, this leads to (univariate) Jeffreys' prior. Problems with many parameter are shown to reduce to a sequential application of the one parameter algorithm (and this does *not* lead to multivariate Jeffreys' rule). For key developments in the theory of reference analysis, see Bernardo (1979b) and Berger and Bernardo (1992); for a simple introduction see Bernardo and Ramón (1998).

14. What is the role in theory and in practice of upper and lower probabilities? Upper and lower probabilities have been important players in the theoretical search for descriptions of uncertainty which might provide an alternative (or a generalization) to the use of probability theory for this purpose. For instance, proponents of 'knowledge-based expert systems' have argued that (Bayesian) probabilistic reasoning is incapable of analyzing the loosely structured spaces they work with, and that novel forms of quantitative representations of uncertainty are required. However, alternative proposals, which include 'fuzzy logic', 'belief functions' and 'confirmation theory' are, for the most part, rather *ad hoc* and have so far failed to provide a general alternative. For some interesting discussion on this topic, see Lauritzen and Spiegelhalter (1988).

Any acceptable approach to statistical inference should be quantitatively coherent. The question of whether quantitative coherence should be precise or allowed to be imprecise is certainly an open, debatable one. We note, however, that it is possible to formalize imprecision within the Bayesian paradigm by simultaneously considering all probabilities compatible with accepted assumptions. This 'robust Bayesian' approach is reviewed in Berger (1994).

REFERENCES

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

Berger, J. O. (1994). A review of recent developments in robust Bayesian analysis. *Test* **3**, 5–124., (with discussion).

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229–263.

Bernardo, J. M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. and Juárez, M. (2003). Intrinsic Estimation. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford: University Press, (to appear).

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 1–35.

Bernardo, J. M. and Rueda, R. (2002). Bayesian Hypothesis Testing: A Reference Approach. *Internat. Statist. Rev.* **70**, (to appear).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Berry, S. M. and Kadane J. B. (1997). Optimal Bayesian randomization. *J. Roy. Statist. Soc. B* **59**, 813–819.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68. Reprinted in 1980 as 'Foresight; its logical laws, its subjective sources' in *Studies in Subjective Probability* (H. E. Kyburg and H. E Smokler, eds.). New York: Dover, 93–158.

Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker eds.). Dordrecht: Reidel, 175–257 (with discussion).

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.

Kardaun, O. J., Salomé, D., Schaafsma, A. G. M., Willems, J. C. and Cox, D. R. (2002). Reflections on fourteen cryptic issues concerning the nature of statistical inference, *Internat. Statist. Rev.* **70**, (in this issue).

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.

Kuhn, T. S. (1962). *The Structure of Scientific Revolution*. Chicago: Phoenix Books.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures, and their application to expert systems. *J. Roy. Statist. Soc. B* **50**, 157–224 (with discussion).

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley. Second ed. in 1972, New York: Dover.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656. Reprinted in *The Mathematical Theory of Communication* (Shannon, C. E. and Weaver, W., 1949). Urbana, IL.: Univ. Illinois Press.

Stone, M. (1969). The role of experimental randomization in Bayesian statistics: Finite sampling and two Bayesians. *Biometrika* **56**, 681–683.