International Statistical Review (2003), **71,** 2, 277–318, Printed in The Netherlands

Reflections on Fourteen Cryptic Issues Concerning the Nature of Statistical Inference*

O.J.W.F. Kardaun[1], D. Salomé[2], W. Schaafsma[2], A.G.M. Steerneman[2], J.C. Willems[2,4] and D.R. Cox[3]

[1] *MPI für Plasmaphysik, Garching, Germany.* [2] *Groningen University, The Netherlands.* [3] *Nuffield College, Oxford, UK.* [4] *University of Leuven, Belgium.*

Knowledge can be communicated by teaching, . . . , but all teaching starts from facts previously known, as we state in the Analytica Posteriori, since it proceeds either by way of induction, or else by way of deduction.

Aristotle, Ethica Nicomachia VI. iii. 3.

Summary

The present paper provides the original formulation and a joint response of a group of statistically trained scientists to fourteen cryptic issues for discussion, which were handed out to the public by Professor Dr. D.R. Cox after his Bernoulli Lecture 1997 at Groningen University.

Key words: Bayesian analysis; Decision analysis; Distributional inference; Epistemic uncertainty; Foundations of probability theory and statistics; Likelihood approach; Nonparametric statistics; Objectivistic and personalistic interpretation of probability; Possibility theory; Prior distributions; Semiparametric statistics; Time series analysis.

Introduction

On the occasion of the (Johann) Bernoulli Lecture 1997 at Groningen University on The Nature of Statistical Inference, the last author handed out fourteen cryptic issues for discussion to the public. The original formulation of these issues is presented as well as the answers provided by a group of scientists after internal discussion.

1. *How is overconditioning to be avoided?*
2. *How convincing is A. Birnbaum's argument that likelihood functions from different experiments that happen to be proportional should be treated identically* (*the so-called strong likelihood principle*)*?*
3. *What is the role of probability in formulating models for systems, such as economic time series, where even hypothetical repetition is hard to envisage?*

*Part of this work was funded by a collaboration between IPP and Euratom. The contents of this work is the sole responsibility of the authors. In particular, the views expressed therein are not to be construed as being official and do not necessarily reflect those of the European Commission or the Max-Planck-Gesellschaft.

4. *Should nonparametric and semiparametric formulations be forced into a likelihood-based framework?*
5. *Is it fruitful to treat inference and decision analysis somewhat separately?*
6. *How possible and fruitful is it to treat quantitatively uncertainty not derived from statistical variability?*
7. *Are all sensible probabilities ultimately frequency based?*
8. *Was R.A. Fisher right to deride axiomatic formulations* (*in statistics*)*?*
9. *How can the randomisation theory of experimental design and survey sampling best be accommodated within broader statistical theory?*
10. *Is the formulation of personalistic probability by De Finetti and Savage the wrong way round? It puts betting behaviour first and belief to be determined from that.*
11. *How useful is a personalistic theory as a base for public discussion?*
12. *In a Bayesian formulation should priors constructed retrospectively, after seeing the data, be treated distinctively?*
13. *Is the only sound 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?*
14. *What is the role in theory and in practice of upper and lower probabilities?*

These cryptic issues nailed at the door of the Aula of Groningen University provided a nagging challenge to many people, the poser of the questions not excluded. Two groups of students wrote to Professor Cox who kindly responded. Several tentative individual reactions were given and issued as internal reports [36, 75]. It was decided to combine efforts and formulate a well-discussed reaction as a group. Through L.J. Smid [142, 125], who taught at Groningen University, some members of the group were influenced by D. van Dantzig [60] and the Significa Movement [93]. This movement was inspired by the ideas of Victoria Welby [148] and, more recently, has been reported to have anticipated Austin's 'speech acts' [6, 8]. The people of the Significa Movement thought that making precise definitions would help to resolve controversies. Their positivistic attitude is sometimes described as that of the modernists, in statistics exemplified by R.A. Fisher and K. Pearson.

The role of the last author (D.R.C.) needs some comments. After having devised the original questions, he has been involved in correspondence to clarify the issues and to deal with some of the points that did arise. The paper represents primarily what can be called 'a Groningen view of statistics'. While D.R.C. is very sympathetic to many of the answers given, there are inevitably differences, mostly of shades of emphasis but sometimes of more substance. It has seemed most fruitful to leave the paper as a statement of the view of a distinctive school and not to attempt to resolve the points of difference in every detail.

This paper is an attempt to formulate some sort of *communis opinio* among several specialists, with a common as well as a somewhat different background, about the issues formulated. In the view of the authors, a purely subjectivistic and personalistic approach to the world around us is not of much value. A purely objectivistic approach, on the other hand, is also not entirely satisfactory, since it fails to recognise explicitly the role of judgement at various stages and interpretation and typically deals only with uncertainty arising from a certain kind of 'statistical' variation. Even within this framework, it does not provide compelling statistical solutions: if it establishes a procedure to be 'optimal', it does so only by simplifying the context. Hence, an eclectic attitude is preferred, even if the danger is realised that this can blur the underlying issues. Perhaps there are even many situations where a statistician's proposed recommendation depends so much on the choices he makes that he should rather return the problem to its owner without being explicit about a solution. The role of statistical theory and indeed of the individual statistician is to provide a broadly acceptable framework of concepts and methods to aid the analysis and interpretation of data and, in suitable cases, to provide a basis for decision making. The public acceptability of these ideas as a base for

communication is important. Therefore the freedom to choose *ad libitum* from the various meta-statistical 'schools' has its limitations, since a 'professional' answer is required. Everybody can make statements about tomorrow's weather, a patient's disease, etc., but it is the meteorologist or the pathologist whose statements we should respect: they are the experts who are engaged to codify their behaviour, to calibrate their opinions, to discuss basic principles, and thus form well-balanced conclusions that are to be considered as 'scientifically sound'. They are the professionals in their domain. We, applied and theoretical statisticians, try to be professional as well. The fact that the subject matter does not admit unicity of solutions does not delimit personal responsibility. To the contrary, it implies that personal (or group) responsibilities are involved in deciding whether or not a solution should be imposed and, if so, which solution should be recommended. In this respect the statistician's work resembles that of a pathologist for instance. What we can do however, in the absence of unique solutions, is to aim at some form of *intersubjectivity* by combining opinions and by sharing earlier experiences. In statistical science we try to come to peace with 'Pearson', 'Fisher', 'Neyman', 'Jeffreys', 'Savage', and other scientists who were motivated by real-world issues and have in common that they thought deeply about methods of making inferences on the basis of data in conjunction with models containing unknown parameters. The metaphysical and epistemological character of the situation excludes the possibility of *compellingness* of the inferences or *infallibilism* of their maker. Popper has made the statement that *induction is a myth*. He was referring to situations where the first n elements of a sequence had to be used to make a statement about the elements not yet seen. We share his doubts. On the other hand, as statisticians we know that *induction is a must*. The statistical character of "a sample of n elements from some population" and the probabilistic idea of "independent repetition" are the paradigms that show that induction is not always a myth though, of course, it will always be a matter of approximation: epist$\bar{e}$m$\bar{e}$ (true knowledge) will neither be possible on the basis of a sample nor on the basis of the first n elements of a sequence.

Statisticians are usually more fascinated by the concrete paradigm of a sample from some population (for instance patients in a hospital) than by the mathematical paradigm of a sequence of independent repetitions. It is quite natural for mathematically oriented probabilists to favour independent repetitions of some random experiment because this allows the establishment of nice mathematical theorems, such as the laws of large numbers and the Glivenko–Cantelli theorem. In Loève's book [92] this mathematical approach to the problems of statistics is expressed by referring to Glivenko–Cantelli as *The Main Statistical Theorem*. This probabilistic attitude is somewhat alien to that expressed by the schools of thought attached to the names of Fisher, Neyman, Jeffreys, and Savage, though, of course, none of these will seriously belittle the importance of asymptotic theory or of establishing mathematical theorems. The only difference is that the four schools indicated pay more attention to 'exact' small-sample inference, which is more adapted to the situation where *a given set of data* has to be evaluated. The real difference between the group indicated by the names of Fisher, Neyman, Jeffreys and Savage, and that indicated by the names of Borel, Kolmogorov and Loève, is, perhaps, less pronounced than just suggested. Both will agree that induction is a myth as well as a must. The group Fisher *et al.* will emphasize that mathematics has a task in developing appropriate intuitions for exact small-sample inference. The group Borel *et al.* regards this task as a "mission impossible". They, of course, have their own intuitions which they can share with others or derive from work of the first group. Anyway, both groups have in common that they try to explore the probabilistic context as well as they can.

There are many situations, cryptic issue 3 is referring to some of these, where the 'stochasticity' and 'linearity' assumptions incorporated in a statistical model have their origin rather in the mind of the data analyst than in the real world. One of us (JW) has a background in Mathematical Systems Theory [105]. While being sceptic about such assumptions he responded to cryptic issue 3. The other authors share many of his concerns, even though some of them (in particular AS) participated in the analysis of a multiple time series representing the Dutch economy. This analysis led to predictive

statements about features of next-year's Dutch economy in the form of an estimate $\pm$ standard error. Evaluation afterwards taught that the standard errors were not meaningless. In [41] a table is presented with actual outcomes, predictions and standard errors of the national consumption growth from 1981 to 1990. The ratios between the absolute error and the standard error were

$$1.64,\ 2.18,\ 0.95,\ 0.34,\ 0.21,\ 0.20,\ 0.39,\ 0.49,\ 0.0,\ 0.54.$$

They resemble a sample of the absolute value of a standard-normal variable as well as a sample of a log-normal distribution, $\log X \sim N(-0.59, 0.84)$. In the latter case, the observation 0.0, actually 0.006 [41], seems to be an outlier. Another area of investigation to which issue 3 bears is the scaling, based on present-day experiments, of the confinement time (a numerical expression of the quality of thermal isolation), related to the prediction of the performance ('fusion power yield') [79] of heated plasmas in a prospective magnetic fusion device [77, 11, 131, 10], oriented towards a future source of energy [101]. For this topic, the reader is referred to [152, 27, 72, 106, 76, 77, 78, 87]. Related to this is the question whether or not a particular type of confinement regime is reached at the reference operating point [72, 73, 117, 77, 133], and the prediction of plasma temperature and density profiles [71, 94, 77, 97]. Some statistical areas the authors have been involved in are variable selection and model screening, discriminant analysis, growth curves, multivariate analysis, survival analysis, meta-analysis, design of experiments, size and shape analysis, distributional inference and other foundational issues, such as 'structuring the inferential contest' [41, 3].

After this introduction we will now turn to a joint response to the cryptic issues formulated by the last author. Since they are intended to stimulate further investigation by discussion, we do suggest the reader to put the article away for a moment and to try to answer these questions—even if provisionally—by himself, before continuing to read and to contrast his findings with ours.

1 How is Overconditioning to be Avoided?

If the phenomenon of overconditioning is restricted to its most primitive form then the question corresponds to *the problem of the reference class* formulated by Reichenbach [113], after his PhD thesis [112]. In the section about the 'Frequency Interpretation of the Probability of the Single Case' he stipulates that we should consider *the narrowest class for which reliable statistics can be compiled.* That this qualitative statement does not really settle the issue, is apparent from [48, 86, 145] and is illustrated by the following example, which was presented by the last author during his Bernoulli Lecture. Assume that you have an urn with one hundred balls numbered $1 - 100$. Some are black and the others are white. A random sample of 50 taken without replacement shows 25 black balls. Because of the random sampling there is no difficulty in inferring that of the 50 balls unobserved, roughly 25 are black, and a confidence interval could be calculated if required.

Now suppose that you have the following additional information. Some balls are made of wood, some of steel, the rest of copper. Some balls are made in Groningen, some in Amsterdam, and the other ones in Eindhoven. Some are small, some are medium, the rest is large. Some are old, some are of medium age, and the rest is new, and so on. Also, the person who painted the balls is known to have treated different kinds of balls differently. For example, he or she may have painted black all the large balls from Amsterdam that are new, etc. However, what he or she actually did is totally unknown. Now consider the following situation. Suppose there are a few (exact number unknown) in the un-sampled set from Eindhoven, of copper, and old, but none of these have been sampled. Therefore there is not any empirical information about this particular group of balls. This influences the probability of blackness of a ball sampled from Groningen. The magnitude of this influence is unknown. Therefore, by conditioning over all conceivably relevant information, one can in the end still not construct a reliable interval estimate of the probability to draw (in future) a black ball from Groningen, in view of the (unknown) fractions of unsampled balls in the population. This is

an extreme instance of a difficulty that you would get into by over-conditioning. The simplicity of the above formulation should not disguise the fact that this is one of the most difficult of the cryptic issues formulated. Overconditioning has its counterpart in oversimplification. Hence it is a Scylla–Charybdis situation. It may happen that sailing between these monsters is so dangerous and the reward, in terms of accomplishment, is so small that the statistician had better return the problem to its owner and wait for better weather.

Overconditioning is not an isolated subject. It is related to incorporating too many variables in an analysis and to overfitting, i.e., using models with too many parameters. Sometimes the use of too many variables or parameters manifests itself in the numerical analysis, e.g. when matrices become ill-conditioned. The real issue, however, is much deeper and has a statistical nature. In his letter to two econometrics students (M. Timmermans and K.J. Veltink) the last author discussed the phenomenon of overfitting. He drew a distinction between the fitting of purely empirical models and of those where some reasonably established theory is involved, where it may be wise to include explanatory terms even though their precision is rather poorly evaluated. The other authors wholeheartedly agree though they feel that the prevalence of models based on 'reasonably established' theory should not be overestimated: many situations from practice are so complicated that opinions diverge because of the fact that a number of different theories and models exist, each of which is only in theory reasonably well-established. Even in experimental physics the models fitted to the data are often largely constructions in the mind of a group of investigators rather than the result of already empirically well established theory. This occurs especially in practically complicated situations where the avoidance of oversimplification has resulted in rather complicated models and one needs to worry about ill-conditioning and overfitting.

If one tries to go to the bottom of this cryptic issue then there are various levels for the discussion:

(1) the *fundamental* level where one tries, in accordance with the ideals of the Significa Movement, to specify the meaning of words like "overconditioning" and "overfitting", and the lack of performance these niceties generate (lack of calibration, display of overconfidence, lack of accuracy and reliability). Briefly speaking one may state that overconditioning reduces too strongly the size of the reference class by taking too many (discrete) variables into account. Overfitting means choosing a member from too rich a model class or, equivalently, adapting too flexible a model to the available data. This usually leads to unreliable extrapolations and sometimes even to unreliable interpolations. Estimating equations, e.g. based on least squares or maximum likelihood, of the model parameters will be ill-conditioned in the sense that their (unknown) confidence regions tend to be very large in some directions. (These confidence regions are unknown and can only inaccurately be estimated.) At least in linear models one can geometrically see that such a situation occurs if in the space of the regression variables ($\mathbb{R}^p$ for p variables) the range of the available data is not large with respect to the actually possible measurement errors, for simplicity modelled as random ones. This leads to an unstable behaviour of some directional derivatives of the regression plane, i.e. of certain linear combinations of the regression coefficients. Notably ordinary least squares regression yields in that situation biased point estimates and inaccurate (often too narrow) interval estimates. (It is noted that methods based on errors-in-variable models may provide here some relief, at least if the ill-conditioning is not too strong.)

(2) the *pragmatic* level where specific proposals are made to avoid overconditioning (and overfitting), for instance by testing in a two-group discriminant analysis context the null hypothesis that the population Mahalanobis distance is not increased by incorporating a number of additional variables, see [108]. If such a null hypothesis is not rejected, then one may proceed on the basis of the less complex model where these additional variables are simply ignored. This testing-of-null-hypotheses approach is considerably appealing because of the clearness of its prescription. There are some difficulties, however. Which additional variables or parameters should be chosen to govern the alternative

hypotheses? Which significance level should be chosen? At an elementary multiple-regression level: should we delete explanatory variables if the corresponding regression coefficients are not significantly different from 0? In his letter to the two econometrics students, the last author formulated his viewpoint as follows: 'it may be wise to include explanatory terms even though their precision is rather poorly evaluated'. This is in line with an experience of Rao [109] with a data-based discriminant function which displayed a very good discriminatory performance in an evaluation set, though the coefficients were extremely unreliable. An attempt to improve the discriminatory performance by inserting 0 for the insignificant or illogical discriminant function coefficients (for some of them the sign was wrong) was counterproductive: instead of an increase of performance a substantial decrease was observed. Such experiences (and the fact that they are communicated) are extremely important to develop appropriate intuitions.

A more parsimonious model formulation may damp out the effect of ill-conditioning from overfitting. Such a model formulation may for instance be based on testing hypotheses as discussed above, or alternatively by introducing 'stiffness', with respect to some predetermined value(s) of the directional derivative(s), in the direction(s) where the data variation is scarce. This does not always help however, especially not if the dependence of the response variable in those ill-determined directions is of special interest for the investigation. Amelioration is in such a situation possible by using estimation techniques based on errors-in-variable models, see [5, 62, 28], and from the experimental side by either reducing the measurement errors in the regression variables or by extending the range of the data in the direction(s) where it is most needed.

(3) the *mathematical level* where one tries to specify exactly how to sail between this Scylla and Charybdis such that, on the average, certain characteristics of performance are most satisfactory. Such attempts are, perhaps, counterproductive from a practical point of view. In the context of variable selection in discriminant analysis we refer to [111], especially because its original motivation came from the physical anthropologist G.N. van Vark whose computations (in the early seventies) suggested the existence of an 'optimal' number of variables in a discriminant analysis. Incorporating more variables resulted in a decrease of discriminatory performance.

2 How Convincing is A. Birnbaum's Argument that Likelihood Functions from Different Experiments that Happen to be Proportional Should be Treated Identically (the So-Called Strong Likelihood Principle) ?

In objectivistic parametric statistics (especially the Neyman–Pearson interpretation) the likelihood ratio is a quantity of primary interest. In subjectivistic statistics (Jeffreys as well as De Finetti), posterior densities, based on proper or improper priors, are of intrinsic interest. Multiplying the likelihood with a constant does not influence the transition, by Bayes's theorem, from prior to posterior probabilities. From this point of view the (strong) likelihood principle (as formulated in the question) seems reasonable. Nevertheless, we have good reasons to reject this principle since, for instance, the context of the experimental design is ignored. If, for example, 3 successes are observed in 12 Bernoulli experiments (a situation discussed in Berger–Wolpert [20]) it makes a difference whether the total number of experiments (12) or the total number of successes (3) is fixed in advance. Nevertheless, the likelihood is proportional in the two cases. Bayesians are sometimes suggesting that, in their view, it is a merit of their approach that the posterior does not depend on the design [44]. It is noted, however, that especially in the case of absence of prior information, the specification of the design may provide relevant additional information about the concrete situation. Since the question refers to Birnbaum's argument, we concentrate the attention to the argument itself. While presenting here our own account to this question, we will not enter into the more extended discussion presented in [68]. As described in [20] the strong (or formal) likelihood principle follows from

(1) the *weak conditionality principle* which can be regarded as a formalisation of an example by the last author referring to the measurement of a location parameter θ either by device 1 or by device 2: the evidence about θ from the combined experiment (choosing the device at random and performing the measurement with the device chosen) is just the evidence from the experiment actually performed,
(2) the *weak sufficiency principle*: if $T = t(X)$ is a sufficient statistic for θ then $t(x_1) = t(x_2)$ implies that the evidence about θ, involved by the observations x_1 and x_2, respectively, is the same.

Before continuing we like to express our gratitude that such principles exist. We have nothing against likelihood inference, conditional inference or the reduction by sufficiency, provided that deviations are tolerated, and dogmatism is avoided.

What is wrong with the likelihood principle? Statistical inferences are mixtures of facts (the data x actually available) and fictions (the model $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$, even though it has a factual core and other contextual ingredients). Now suppose that experiment h designed by statistician (or scientist) h resulted in data x_h which, according to the fictions statistician h decided to incorporate, provided a likelihood function $l_{x_h}(\theta)(h = 1, 2)$. Suppose that these likelihood functions happen to be proportional. Does this imply that the two statisticians should come up with the same inferences? Obviously not. A reasonable suggestion would be that the statisticians should combine the information they have, discuss the reliability of the functions involved, and try to arrive at some joint opinion, for instance in the form of a confidence interval or a distributional inference. This type of approach, however, is beyond the scope of the formal likelihood principle.

As we reject the likelihood principle, we shall have to reject at least one of the two principles Berger–Wolpert [20] used as its basis. In fact, we shall reject both principles.

What is wrong with the weak conditionality principle? Almost nothing, except for the fact that the word "evidence" is not well defined. Berger and Wolpert suggest a very general definition including "any standard measure of evidence, or something entirely new". The last author has used the example to criticise unconditional methods of inference induced by a Neyman–Pearson formulation, see [29]. The example can also be used in the opposite direction. If one has to test $H_0 : \theta = 0$ against $H_1 : \theta = 1$ at a prescribed level α, say $\alpha = 0.05$, then the optimal level-α test is not conditionally (given the device used) of level α. This might be regarded as an argument against the Neyman–Pearson–Wald testing theory, but a more reasonable reaction is that it is wrong to use the weak conditionality principle as a dogma.

What is wrong with the weak sufficiency principle? Almost nothing, except for the fact that it is less compelling than it would seem at first sight. In the Neyman–Pearson–Wald approach it is established that the reduction by sufficiency does not affect risk-function properties if randomisation is allowed (and the loss does not depend on latent outcomes of random variables: in a predictive context one has to be careful). The weak sufficiency principle refers to a mysterious but important concept of evidence. If this is made operational then one should expect that some form of randomisation will be involved. This affects the idea that $t(x_1) = t(x_2) \Rightarrow$ evidence carried by x_1 = evidence carried by x_2, unless the same randomisation is involved. That the reduction by sufficiency is less obvious than is often suggested becomes clear by studying Scheffé's solution to the Behrens–Fisher problem, see e.g. [116]. This solution provides exact confidence intervals but violates the weak sufficiency principle. We can appreciate the exactness of the Scheffé result and, hence, use this to criticise the weak sufficiency principle. The opposite is also possible: criticise the Scheffé solution because it violates the weak sufficiency principle. The Scheffé solution can be replaced by a randomised version. The 'exactness' will then be retained. A further Rao–Blackwell type of improvement is possible such that the weak sufficiency principle is complied with, but this affects the exactness.

Summary. We can enjoy the existence of many principles such as the likelihood principle, the conditionality principle, and the sufficiency principle in particular, and make use of these principles,

even if we are critical with respect to their compellingness.

Exact small sample inference is a mission impossible since compelling solutions do not exist and one always has to sail between Scylla and Charybdis. The principles we use may look very reasonable but there is always some extrapolation involved which, if disregarded, may lead to 'a fallacy of misplaced concreteness' (a concept introduced by Whitehead, see [150]). Induction is a must but compellingness of results, logical validity, is a myth. It would be wrong, however, to belittle such principles. They are the guidelines we need to structure our discussion which, in a sense, is about the quantitative specification of meta-physical considerations. The connotation of the word epistemology is less negative (to mathematicians and physicists) than the word metaphysics. Yet we should not betray ourselves: the work of statisticians is in these nasty areas which are full of fallacies, rationalisations which become misleading if the extrapolations they involve are too bold.

In the present context we distinguish between three fallacies of misplaced concreteness:
(1) the idea of Fisher, see e.g. [49], to use the class $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ as the basis of the statistical analysis; even if the assumption $P \in \mathcal{P}$ is acceptable we have the difficulty that $\mathcal{P}$ is a peculiar mathematical 'extension' of $\{P\}$, needed because P itself is unknown to us, but peculiar in the sense that only P is factual,
(2) the idea of De Finetti [40, 13] to subject epistemic probabilities to axioms similar to those of Kolmogorov's theory,
(3) the idea expressed by the three metaphysical principles occurring in Birnbaum's argument which, stripped to its essentials, states that induction is a *must* and, hence, methods "should" be developed prescribing the inferences one "should" make. Next one continues as if such methods exist. This is the fallacy behind the sufficiency principle and behind reductions by invariance. One discusses the possible outcomes x_1 and x_2 related by a common value of the sufficient statistic or by some transformation. Next one requires that the statistician who observed x_1 should arrive at the same conclusion as the statistician who observed x_2. We are attracted by the lure of appealing simplicity behind such requirements and yet we feel that it is a matter of wishful thinking.

3 What is the Role of Probability in Formulating Models for Systems, Such as Economic Time Series, Where Even Hypothetical Repetition is Hard to Envisage?

Some elements of the first *problématique*, the role of probability in modelling systems from data, are very prevalent in the area called system identification. We will address this issue in this context. System identification is primarily aimed at modelling industrial processes. However, the methods used are very similar to what is done in econometrics.

In order to fix ideas in our remarks about system identification, one can think of two examples from our daily experience. Assume that measurements are taken of the height of the road surface and the vertical motion of a person in the driver's seat of a car moving at fixed forward speed. How would one use these measurements in order to come up with a dynamic relation, for instance in the form of a differential equation, connecting these two variables? As the second example, assume that measurements are taken of the outside temperature and of the temperature in the living room of a house, and that we aim at using these measurements in order to obtain a dynamic law relating these two temperatures.

The procedure that is typically followed in system identification is to choose a model class for relating these variables, say, an ARMAX model (which stands for an auto-regressive moving-average model with additional input variables). Thus some of the measured variables are considered to be exogenous and some endogenous. Their linear dynamic relation is corrected by stochastic terms, whose time-structure is modelled through the MA-part. (With *stochastic* we mean variables with a probabilistic interpretation. Of course, here we meet the difficulty that different schools have a different interpretation of this interpretation, but we will kindly gloss over this.) Subsequently, statis-

tical procedures are invoked to estimate the parameters of this ARMAX model, whose order is kept reasonably low in order to avoid, among other things, over-fitting. Identification algorithms [91], as implemented for instance by MATLAB's System Identification Toolbox (developed by Lennart Ljung from Linköping University in Sweden) or by SAS in the module Econometric Time Series (ETS) [120], provide such procedures. Using sophisticated ideas involving statistical estimation, frequency-domain thinking, approximation, and recursive optimisation, a model is computed. The central idea (although there are other possibilities) is to choose the parameters of the ARMAX model such that the sum of the squares of the one-step ahead prediction errors is minimised. Recent refinements involving for example Neural Networks incorporate nonlinear models in this setting as well.

There is no doubt that this work is among the most important achievements of systems theory. Moreover, because of its immediate relevance to modelling, the all-important issue in applying mathematics, this work is highly relevant for industrial applications. Nevertheless, it is appropriate to ask questions about the relationship between the identified (or estimated) stochastic model and the real system that it tries to capture.

When one faces the problem of fitting a model to an observed vector time series by means of, say, a linear time-invariant model of moderate order, then the difficulty occurs that none of these models will fit the data exactly. How should one proceed? In ARMAX models, we choose to assume that in addition to the exogenous inputs, there are some additional (unobserved) inputs, let us call them latent inputs. These latent inputs are assumed to drive the system also, and to influence the measurements (often additively). The next step is to assume that these latent variables form stochastic processes [30]. However, in the applications that we have in mind, the systems are simply (predominantly) deterministic, see [127, 105]. If, in the car example, we drive an identical car over the same road, then we will roughly see about the same motion for the driver. Also, it is unlikely that they are the much advertised measurement errors that justify the introduction of stochastic aspects in the model. Most modern measurement devices are very precise and their inaccuracies are for all practical purposes negligible. If there are deviations between the sensor output and the intended measurement, then these are more likely due to nonlinear and dynamic effects than to the pick-up of random disturbances. It is hard to envision sensors that process measurements in a *signal plus additive noise* fashion.

These and other considerations lead to the conclusion that in most identification problems the lack of fit between the deterministic part of the model and the measured data will be due to *approximations* [61] and not to the presence of *random inputs*. These approximations will be due to such things as the use of linear models for nonlinear systems, of low order models for high order systems, of time-invariant models for time-varying systems, and of neglecting unmeasured inputs. The question is whether it is reasonable to signal these approximations through stochastic terms in the model.

Statistical system identification is an effective way to deduce models from data. The deterministic part of the model can be expected to provide a reasonable idea of the relation between the exogenous inputs and the endogenous outputs. The stochastic part of the model provides a reasonable idea of how to assess the misfit between the data and this relation. Moreover, since the stochastic part is dynamic (MA), it also subtly provides this misfit with memory, which is a very reasonable thing to do. Thus as an approximation, the deterministic part of the model gives us a sort of optimal (weighted) least squares fit, while the stochastic term tells us how to assess the misfit. Thus, all is well as long as one does not become mesmerised by the interpretation of the stochastic terms as providing a measure of the relative frequency of future events, of percentages of probabilities that certain things will happen, or of degree of belief in a model one doesn't believe anyway. As a data reduction algorithm, summarising the observations, the ARMAX model is very effective. As a model of reality, it is more difficult to justify. Issues as falsification, etc., are philosophically hardly relevant: the model is not meant to be more than an approximation. The situation is different and needs even

more care if the individual coefficients and their inclusion or exclusion is of intrinsic subject-matter interest or if certain individual parameters are to be given a provisional causal interpretation [34].

A related question that results from this is the sense of studying questions as *consistency* and *asymptotic efficiency* in system identification. We offer but one thought on this. When an ARMAX model is fitted to a time series, then it is logical to demand that the resulting fit will as quickly as possible become *exact* (or *optimal*) in the purely theoretical case that the time series happens to be produced by an element of the model class. In other words, *algorithms should work well with simulated data, which are modelled in such a way that they rather accurately mimic the corresponding actual measurements*. While we view this as the only possible justification of much of the theoretical work that goes on in identification theory, we have ambivalent feelings on the value of this justification when specifically applied to the practice of ARMAX modelling. On the one hand, it seems reasonable and it will guide the setting up and the evaluation of algorithms. It has certainly been the topic of numerous journal articles. On the other hand, consistency, for example appears a somewhat irrelevant issue when one consciously decides to fit a finite-dimensional linear time-invariant stochastic model to an infinite-dimensional nonlinear slowly time-varying deterministic reality.

The premise: *algorithms should work well with simulated data* seems to be the only feasible paradigm for justifying the algorithms used in system identification (and in adaptive control). Perhaps it is even the only way to think about assessing such algorithms scientifically (although it is good to keep an open mind about alternative philosophical approaches). This implies one should appreciate the importance of much of the theoretical work that goes on in system identification (including algorithms that are based on the idea that one should start by searching for the most powerful unfalsified model). But there are two caveats: first, what class of 'simulated data' should one consider as the testbed, and, secondly, what does 'work well' mean? And these caveats, particularly the first one, are all too easily glossed over. Why is it reasonable, when fitting a very simple model to a very complex reality, to test an algorithm against stochastically generated data, and why is it reasonable to take consistency as the first requirement for good performance?

The above considerations are very relevant to the role probability has to play in modelling econometric time series. For interesting monographs on time series analysis we refer to [30, 102, 147, 39]. In econometrics, we are definitely fitting a simple model to a complex reality. To interpret the lack of fit stochastically should not be done, unless very very cautiously.

But perhaps there are other origins of stochastic randomness in econometrics. Of course econometric data are influenced by more inputs than are taken into account in the model while the latent inputs are to some extent real. (Of course, the same is true for our car example: the wind is also an input, the driver may be nervous and not hold the speed constant, the fuel quality may be variable, etc.) There are things as unmodelled political events, swings in consumer confidence, technological innovations, the euphoria of the financial markets, and what have you. Maybe it are these effects, much more than the lack of fit, that account for the stochastic part of the model. Perhaps, but we doubt it. But even if "perhaps" becomes "surely", the question remains why these latent inputs should be modelled using probability and statistics. Why does the sum of technological innovations, political events, and consumer mood behave like random variables?

We now turn to the final issue, the impossibility of even hypothetical repetition. Of course, it is usually impossible to experiment in macro-econometric systems and repeated observations under identical circumstances are difficult to conceive. However, is this really the difficulty in econometric modelling? Let us speculate a bit. Imagine that in 100 worlds (or countries) we would have observed for a large number of consecutive years identical economic time series. Is it not more reasonable to expect that the next year we will observe about the same value, than to expect that we will observe a scatter due to stochastic effects? It is obviously unclear why we would see a scatter and hence repetition, even if it were possible, may teach us very little new. Of course, it may still be difficult to predict the next event, and different theories, different statistical modelling techniques, different

political insights, and different sayers will come up with different predictions. But that is due to ignorance, not to an intrinsic scatter. However, all this is mere speculation.

On the other hand, is it really impossible to experiment in economics? Certainly, in the sense that it is impossible to do "controlled" experiments. However, "nature" repeatedly experiments for us: each quarter and in each country a new experiment is performed. The question seems more why these experiments are not adequate in order to come up with a good model, than that experiments are not done. Of course, economists will object that the UK economy should not be considered as an experiment on the cherished Polder-model, and that the Dutch economy 10 years ago does not yield an experiment for today's. We agree, but, obviously, it does not follow that declaring an event "unique" automatically leads to stochastic models.

In conclusion, there is a good case to be made for statistical system identification algorithms and stochastic models, as providing a good approximation method and a pragmatic way of assessing the lack of fit. However, one should remain cautious not to take all this too literally and not to overinterpret in actual practice concepts like statistical significance, consistency, efficiency, and asymptotic normality.

4 Should Nonparametric and Semiparametric Formulations be Forced into a Likelihood-based Framework?

The likelihood approach, anticipated by Daniel Bernoulli [21], enjoys the qualities of asymptotic efficiency as well as, nowadays, practical computability for large classes of parametric problems. These advantages are severely dimmed if the number of parameters increases proportional to the number of observations. Inconsistency results and identifiability problems (among others for errors-in-variable models) are known since the fifties [82]. This suggests that a likelihood-based framework is inadequate if non- or semiparametric models are considered. Nevertheless various versions of partial likelihood [31, 32, 151] provided asymptotically valid and useful results for certain semiparametric models in the context of survival analysis [31, 69, 83].

Instead of the likelihood function it can sometimes be useful to optimise some other objective function, less dependent on the probabilistic assumptions. In his Theoria Motus [54], Gauss used normality assumptions to generate least squares theory. In his later work [55] he preferred to follow Legendre by simply minimising the sum of squares without postulating normality for the error distribution, see also [137]. Hence, potential functions of which the gradients yield estimating equations that produce consistent estimates, possibly at the loss of some efficiency, are useful anyhow, whether or not they live in an infinite dimensional space, see e.g. [88].

Since the eighties, research interest in general semiparametric models has increased and the theory of likelihood, empirical likelihood and estimating equations has been developed further, see e.g. [90, 43, 95, 114, 107, 59, 18, 141].

Whether semiparametric methods should be forced into such a framework depends on the purpose of the analysis. If the robustness of procedures (against deviations of the specified model assumptions) is an important purpose then, obviously, one will have to go beyond the likelihood framework. Some of us have worked in the area of nonparametric density estimation [25]. The idea was to start from some initial guess of the density. This initial guess may very well be based on a parametric model and, for instance, a maximum likelihood estimate of its parameters. Subsequently, however, a process of nonparametric fine-tuning of the initial guess is carried out which goes beyond a likelihood-based framework. The conclusion is that some, but not all, formulations should be forced into a likelihood-based framework.

5 Is it Fruitful to Treat Inference and Decision Analysis Somewhat Separately?

Our answer is definitely yes. Statistical inference is oriented towards the scientific goal of describing, explaining and predicting Nature's behaviour on the basis of a set of data x. This is largely a

matter of discussion. Statistical inferences can have various forms. The specification of a statistical model $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ is part of it. In decision analysis the discussion is focused on the objective to take a decision or to perform an action on the basis of the data x. This involves a variety of discussions and specifications, e.g. that of perceived loss or utility functions accompanying the various a priori possible actions.

Wald's theory of statistical decision functions [143] established that concepts from decision analysis and, especially, the theory of games offer a perspective for a number of situations from statistical inference. This loss-function perspective provides a basis for comparing methods of inference. In statistical inference, however, loss functions are mainly constructions of the mind. That is why related concepts of unbiasedness, exactness, similarity, etc., cannot be dispensed with. The theory of best unbiased estimators is paradigmatic because it provides methods of inference "valid" for a large family of loss functions. In distributional inference a similar theory can be developed.

After these preliminary reflections we turn to the question formulated, which we divide into three issues:

(1) is it fruitful to develop a theory of statistical inference without specifying aims in terms of future actions, utilities, etc.?
(2) is it fruitful to develop a theory of decision functions concentrating on future actions in general, ignoring the issue of statistical inference?
(3) is it fruitful to develop these theories somewhat separately?

The answer to (1) is yes because the discussion about future actions requires specifications of utility functions, etc., which themselves are subject of a discussion. This discussion would benefit from the existence of preliminary statistical inferences, assessments of probability, interval estimates, distributional inferences, made *independently* from the specific goals and utility considerations of the field of application. The answer to (2) is yes because the issues in deriving the preliminary statistical inferences are different from those in the context of general decision analysis. Hence, in spiritual accordance with Montesquieu's Trias Politica, also our answer to (3) is yes.

Finally, we remark that treating the above two areas of scientific endeavour somewhat distinctly is, among others, also useful to describe the conflict of interest that may arise between the goals of practical decision analysis (e.g. how to treat a particular patient or to optimise the performance of a particular nuclear fusion device) and of statistical inference (how to estimate accurately, and reliably, the physical effect of a certain treatment), see e.g. the introduction of [70, Ch. 15, part 2], which was inspired by [129].

6 How Possible and Fruitful is it to Treat Quantitatively Uncertainty not Derived from Statistical Variability?

Before answering this question, we should first specify the meaning of 'statistical variability'. We understand this expression in the sense of variability due to random fluctuations and, hence, open to probabilistic modelling and statistical analysis.

Usually the statistical analysis does not follow directly from the probabilistic model because this lacks either unicity or practical adequacy. The uncertainty involved, e.g. with respect to the choice of prior in a Bayesian context, is of the same kind as other forms of uncertainty not derived from statistical variability. Probabilistic terminology can be exploited to express such (epistemological) uncertainties. In practice many numerical assessments of probability are obtained by merging facts and fictions. It is then questionable whether probabilistic terminology is the most appropriate one. If the proportion of fiction is large then these probabilities should certainly not be treated as if they were of the type studied in Kolmogorov's theory. In fact, alternative ways to express uncertainty may be exploited. In fuzzy logic, artificial intelligence, mathematical sociology, but also in daily life,

attempts are made to quantify uncertainty, and related concepts such as ambiguity and discord [84]. Some of these approaches are based on plausibility and belief measures rather than on probability measures, see [130, 51]. We have already expressed concern about probabilistic terminology in the answer to cryptic issue 3 and will return to the aspect of the belief measures just mentioned in our answer to cryptic issue 14.

A simple answer to the present question is that it is always possible to express uncertainty by using numbers between 0 and 1. How fruitful it is depends on the situation and on the answer to the question whether scientific purposes are aimed at or other ones. In some situations different research workers, using the same information, make assessments of probability which show almost no agreement. In such cases it is not fruitful, from a scientific viewpoint, to treat uncertainty quantitatively, and it is better to say that "nobody knows". From a non-scientific viewpoint the situation may be different. From his Nicomachian Ethics [7], one may infer that Aristotle would be inclined to state that the prudent ('phronimos') man will adapt his statements of probability to the purposes he has: winning a battle or political dispute, healing or, at least, comforting a patient, etc. In a different context, a pathologist visiting one of us (WS) for statistical consultation, while being confronted with "expressing uncertainty in probability" exclaimed that he did not like this at all: "I do not need such expressions of uncertainty. I never use probabilities since I want my voice be heard".

Hence, it is always possible but the fruitfulness depends, from a scientific point of view, on the agreement among the inferences when different, well educated, statisticians are performing analyses of the data which are available to them all. It will happen very often that consensus exists about extreme cases (hypotheses or patients) whereas the intermediate cases are doubtful, not only in the sense that probabilities are far from 0 and 1 but also in the sense that different experts make different assessments of probability. This leads to the somewhat paradoxical statement that "probabilities are very fruitful if they are close to 0 and 1 but almost useless if they are in the range for which they are intended". This statement does *not* imply that statistical analyses are useless: one will have to do various analyses and make the associated computations before one can decide whether the probabilities are close to 0 and 1 or in the range between.

7 Are All Sensible Probabilities Ultimately Frequency Based?

First, we state that we are in favour of using consistently a word distinct from 'probability' for each type of uncertainty not based on any frequency aspect, for instance, in broad agreement with [36] and extending somewhat [15], 'degree of conviction' for a 'personalistic probability' (De Finetti, Savage), 'verisimilitude' for a personal probability of an ideally coherent and objective person (Jeffreys) and 'credence' or perhaps even 'epistemic probability' for an intersubjective (non-frequentist) probability, a 'reasoned degree of belief' formed by careful scientific investigation, discussion and evaluation of various aspects of a problem. An instance of the latter is provided in [79]. Shafer [130] proposes to use the word 'chance' for objective (aleatory) probability, and the word 'probability' for all kinds of non-objective (epistemic) probabilities. The important point is here that different names should be given to different concepts. This linguistic convention is sensibly adhered to in [130, 84, 145], among others. Of course, *if* it is clear from the context, then, *par abus de langage*, one may use the word probability in all these cases. However, one should not be surprised that this usage leads to crypticism and misunderstanding in those situations that this condition is not fulfilled. Secondly, we remark that the issue bears a cryptic character because of the lack of precise definitions of the words "sensible", "probability", "ultimately", and "frequency-based". Obviously, the answer to the question depends on the clarification of these meanings. Instead of presenting a lengthy analysis of each of these concepts, we just mention that the word probability is used to describe several different notions. Some are mathematically well defined: the probabilities from the games of chance whose numerical specification is a matter of accepting some axiom of equiprobability and the aleatory probabilities

from the theory of probability which are well defined but unknown and yet real numbers and derive their meaning from the theoretical laws of large numbers. Thirdly, we have the assignment of a concrete number between 0 and 1 for any form of epistemic probability. Finally, it is noted that in the usual treatment of quantum mechanics, the notion of probability is used as an intrinsic property of, for instance, electrons and that in statistical mechanics probabilities are expressed in terms of ensembles, which are idealised, hypothetical populations.

In contrast to mathematics-oriented probabilists and theoretical statisticians who want to restrict their attention to ideal situations, epistemologists, many applied statisticians as well as research workers from many other areas, such as medicine and the empirical sciences, usually have little difficulty in accepting the usefulness of "epistemic probabilities", even though they may use other words to indicate such *concrete* numbers between 0 and 1 estimating the (unknown) truth value (either 0 or 1) of an uncertain proposition or event. Such epistemic probabilities are (almost) always mixtures of facts and fictions. Actual data have to be used, as well as a number of theoretical constructions and normative principles. The data can, in many situations, be regarded as frequency based but the fictions are constructions of the mind. A difficulty with such type of probabilities is for instance illustrated by considering the statement that animal experiments show that residential EMF fields, see e.g. [1], are far too weak to cause cancer in humans. Ignoring for the moment epidemiological types of evidence, the question 'what is the probability that this statement is wrong?' bears to the reasonableness of the extrapolation of animal to human experiments. It is certainly arguable that probability cannot usefully be applied quantitatively to such a situation. A personalistic Bayesian proponent has (at any moment) such a probability elicited in principle in the usual way. But how can such a probability be regarded as intersubjectively "reasonable"? Presumably, in part at least, by thinking of roughly similar instances where extrapolation from animal experiments to humans have proved correct, and in that sense thinking, of course in a very approximate way, of a relative frequency in a similar situation.

If "ultimately frequency based" means that observed frequencies are *part* of the specification of the sensible probabilities then the answer to the question is clearly yes. But if "ultimately frequency based" means that *only* observed frequencies and frequency-theoretic considerations (pure probability theory) are involved then the answer is definitely *no*. Some kind of rationalisation or intuition is involved in the specification of the methods of inference which provide the "sensible" probabilities needed. (That such sensible probabilities are not *necessarily* reasonable goes without saying, see for example the discussion of cryptic issue 1.)

8 Was R.A. Fisher Right to Deride Axiomatic Formulations (in Statistics)?

Supposing that Fisher indeed occasionally did so, our answer is of course that he was not entirely right, even though we feel that over-emphasis of axiomatisation can lead to too rigorous simplifications and an ensuing loss of mental flexibility. Axiomatisation is useful (a) as an intellectual game, (b) as a didactically efficient way of summarising a number of different results, while, at the same time, (c) imposing a certain mathematical structure as 'the most appropriate one' for a certain frame of discernment. However, we are inclined to think that he did not deride such formulations *in general*. Although Fisher held definite views, a balanced comparison between the role of deductive reasoning (in probability theory) and inductive reasoning (in statistics) is given in his paper read before the Royal Statistical Society in 1934, see [47]. He sometimes referred to axiomatic theory in mathematics as a source of inspiration. His position is nicely expressed as follows (see [48], Section V):

> "The axiomatic theory of mathematics has not been, and ought not to be, taken very seriously in those branches of the subject in which applications to real situations are in view. For, in applied mathematics, it is unavoidable that new concepts should from time

to time be introduced as the cognate science develops, and any new definition having axiomatic implications is inevitably a threat to the internal consistency of the whole system of axioms into which it is to be incorporated".

We have a lot of sympathy with this phrase, which has given us much food for thought, as well as controversy. It is true that Fisher attacked the method of inverse probability as developed by Jeffreys (see D.A. Lane in [46]) and work of Karl Pearson, sometimes in a deriding way "only a saint could entirely forgive" [124, 154], albeit it is also reported that Fisher had good personal relations with Jeffreys and that, after Fisher succeeded Pearson as Galton Professor at University College in London, 'he would make a point of respectful conversation with "KP" whenever the latter came into the common room for tea' [14]. Fisher did often not appreciate the work of Neyman and Egon Pearson [104], who displayed a 'wooden attitude', and was critical with respect to Wald [143]. Neyman replied with [98] and with [99]. We can now enjoy the polemic style Fisher used, but not without reservation [140]. With some more respect for axiomatisation, i.e. of the precise mathematical specification of the premises in a series of arguments, and for some of his adversaries, Fisher might have perceived more easily the fact that he himself manipulated epistemic probabilities as if they were Kolmogorovian, and, possibly, he would have been less authoritative and more in line with the eclectic attitude many of us display nowadays [37, 110]. But, of course, we would have missed some interesting sentences and some intriguing controversies.

9 How Can the Randomisation Theory of Experimental Design and Survey Sampling Best be Accommodated Within Broader Statistical Theory?

We are not sufficiently familiar with these subjects to provide a satisfactory answer. For technical details related to this question, we refer the reader to [81, 26, 119, 38], and—in the framework of clinical trials—to [134]. While leaving a more adequate answer to the 'younger generation', see [110], we would like to express here solely a concern with respect to the idea that medical investigations should always be performed according to a randomised design. Statistics should be considered as Science's servant, not as its master; more like Gifi, Galton's butler, than like Galton himself. Other aspects than using a randomised design may be more important. For instance, in a medical context: will the patient know which treatment he gets? In the present juncture in which the principle of informed consent is highly valued, an affirmative answer seems often to be regarded as even a moral obligation. However, sometimes, also with a carefully randomised design, this knowledge may prove fateful for the patient. At a certain occasion, one of us (WS) was asked to comment on an (obviously positive) effect of radiotherapy, after surgery, on the (marginal) survival probability of patients with a certain type of brain tumour. During the statistical consultation, the unpleasant suspicion arose that part of the observed effect could have been caused by requests of euthanasia by those who did not receive the therapy: receiving the therapy precluded complying or at least interfered with such requests.

10 Is the Formulation of Personalistic Probability by De Finetti and Savage the Wrong Way Round? It Puts Betting Behaviour First and Belief to be Determined from that.

In first instance, betting behaviour seems to be a reasonable operational way of eliciting, i.e. making explicit to other people (possibly including the betting person himself), personal probabilities. However, the theory of De Finetti and Savage puts betting behaviour before belief and axiomatic theory before betting behaviour.

By requiring that epistemic probabilities satisfy axioms similar to those satisfied by physical probabilities, De Finetti established that coherent behaviour requires the formulation of a prior distribution. (In fact he required that *in addition to* the usual Kolmogorovian axioms certain ex-

changeability properties are satisfied.) The snake in the grass is that there is no compelling reason for epistemic probabilities to behave the same as the physical ones studied in probability theory (usually without specifying their actual values). In our view, it is a *fallacy of misplaced concreteness* to treat these epistemic probabilities as if they were physical. Hence, this is the answer to the question: the personalistic approach of De Finetti and Savage, eloquently advocated by Lindley [89], itself is wrong; it upgrades the status of subjective personal opinion to that of physical evidence. Fisher did something similar when he said that, in the case of absence of information a priori, the making of an observation had the effect that the epistemological status of θ changed from one where nothing is known about it and no probability statement can be made, into one where it has the status of the outcome of a random variable. He failed to note that the specification of such fiducial distribution a posteriori involves the mixing of some fact (the observation) with some fiction. It is easy to criticise the Bayesian approach of De Finetti, Savage and Lindley but such criticism is not fair if nothing is offered to replace this Utopia [50]. The only thing we have to offer is that the derivation of statistical decisions, statistical inferences, distributional ones in particular, etc., should be based on a careful examination of the separate issues. Probabilistic coherency is a nicety which can be aimed at [67], but which should not be pursued indiscriminately. The incoherent behaviour thus displayed [155] will, hopefully, lead to a higher degree of verisimilitude than the coherent behaviour advocated by De Finetti, Savage, Lindley, and others. An analogy is that a system of local linear approximations may be closer to the truth than a single global linear approximation. The theory of eliciting personal degrees of belief has an important spin-off in the concept of *properness* of a utility or loss function [85, 80]. In this respect Savage's Elicitation Paper [123], though difficult to read, is evocative and rewarding.

11 How Useful is a Personalistic Theory as a Base for Public Discussion?

For our critical attitude with respect to the personalistic theory propounded by De Finetti and Savage, we refer to the response on the previous question. Here we will answer the question whether, nevertheless, a personalistic theory can be used as a base for public discussion. The public usually requires statements made by "the profession". This goes beyond personalistic theory restricted to one person. As scientists, statisticians should try to let the data speak. In this respect the De Finetti–Savage approach is not completely useless. A considerable degree of intersubjectivity can be attained by studying a variety of prior distributions and the corresponding posteriors. If the sample size is sufficiently large and the prior distributions are not too weird, then the posterior distributions are quite similar and the public may, perhaps, be impressed by the high degree of inter-statistician agreement displayed. This, however, would be somewhat misleading if no attention is paid to other possibilities of varying the context, e.g. the (parametric) model. Anyway, in a generalised sense, personalistic theory can be useful to some extent as a base for public discussion, especially if a variety of 'persons' is allowed to express their 'opinions' in a probabilistic or possibilistic [130] form. Its use is limited, of course, especially if the strive for coherence and internal consistency is considered to be more important than conformity with the real world [36].

12 In a Bayesian Formulation Should Priors Constructed Retrospectively After Seeing the Data be Treated Distinctively?

This is a cryptic formulation because the meanings of the words "Bayesian" and "distinctively" have to be inferred from the context. We shall answer this question in the context that inferences are made in the form of "probability" distributions. In Groningen this area of research is referred to as "distributional inference" [85, 80, 2]. It contains those parts of Bayesian inference, conditional inference, data analysis, fiducial inference, likelihood inference, predictive inference, structural

inference, etc., where inferences are made in the form of a probability measure. The "probabilities" assigned by such distributional inference (sometimes called inferential distribution) are *epistemic mixtures of facts and fictions*. To organise this somewhat chaotic field we like to use a Neyman–Pearson–Wald approach with proper loss functions and restrictions on the class of procedures. The Bayesian approach has the virtue of providing a convenient framework. It often leads to procedures that are admissible, at least if the loss function is proper, i.e. Bayes-fair [85]. A disadvantage is that the choice of a prior distribution is awkward. Priors constructed retrospectively may lead to inadmissibility in a strict and somewhat narrow decision-theoretic sense. On the other hand, we recognise that there are many ways to construct distributional inferences and to make probability statements. The theoretical inadmissibility of a procedure may be compensated by advantages of unbiasedness, invariance, equivariance, similarity, simplicity, etc. The notion of relevance may be included in this list: it is obviously not an appropriate approach to use a prior which is in conflict with the data. Situations exist where the null hypothesis that such a conflict does not exist can be formulated and tested. If the null hypothesis is rejected and, possibly, in some other cases as well, one would like to adapt the prior before applying Bayes's theorem. Such an "empirical" Bayes approach should neither be considered with distinction nor with disdain. Some suspicion is necessary because the data are used twice, firstly for manipulating the prior and secondly for updating the prior. It is obvious that underlying assumptions of independence (of sources of information) will be violated if the data are, for instance, not randomly divided into two subsets. Much depends on the way the data are used to adapt the prior before using it in Bayes's theorem.

We will usually prefer some other construction of a method of inference. These other constructions also suffer from the fact that they are based on assumptions which, in practice, are usually violated. In this respect, two assumptions deserve special attention: (1) the assumption that the model has been specified a priori (in practice the model will be often selected on the basis of a preliminary inspection of the data), (2) the assumption that no information whatsoever is available a priori about the 'true' value of the parameter. In some instances, a comparative analysis will lead to some intersubjectively acceptable and, hence, "reasonable" and "sensible" procedure. It is also possible that such a comparative analysis indicates that the problem does not allow a sufficiently compelling method of inference. In that case situations may exist where, in spite of the general lack of agreement, the factual data are such that the inferences based on different methods are sufficiently similar for making a dependable statistical statement. Other situations will be such that the case should be dismissed and the problem returned to its owner without making a specific inference, possibly referring to 'the limits of reason' [2]. Statisticians are not hired to guess or suggest, but to make scientifically sound statistical inferences in the light of the data and in spite of some unavoidable uncertainty.

13 Is the Only Sound 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?

The cryptic formulation of this question makes it quite evocative. A semantic analysis is not easy because of the use of a rich choice of modifying words like 'only', 'much', 'rather', 'flat', '(approximate)', 'various', 'viable'. An easy answer to the first question is "No, because distributional inferences themselves (which can be viewed as synthetic judgements a posteriori in the form of a probability distribution, either Bayesian or perhaps non-Bayesian) are a legitimate purpose as well. Similarly, the use of such diffuse, non-personal priors may be helpful in discussing situations where decisions have to be made". Furthermore, Bayesian analysis can play a role in elucidating Stein's phenomenon [135], see also [115], which has a bearing on many applications, among which model selection for catastrophic-type response surfaces [74]. This, obviously, then also answers the second

question. A more profound answer, however, is as follows.

Bayesian work with flat priors seems deceptively similar to likelihood inference. The difference can be made clear by considering what happens under reparameterisation, since the likelihood function does not transform as a probability density. If $l_\theta(\theta; x) = \log L_\theta(\theta; x)$ is the log-likelihood function for a parameter θ, then the likelihood function l_u for any monotonic transformation $u(\theta)$ is given by the direct function composition $l_u(u(\theta); x) = l_\theta(\theta; x)$. In this sense the likelihood function is equivariant under monotonic transformations and so is the parameter location of the maximum likelihood (i.e. the 'likelihood mode') as well as relative parameter locations corresponding to any likelihood ratio, and hence intervals of the type 'full-width at a certain fraction of the maximum'. Obviously, the relation between a confidence interval based on the asymptotic approximation of the likelihood-ratio statistic and, for instance, a somewhat similar interval based on the exact small-sample distribution of this statistic (the latter being possibly complicated but to some extent accessible by stochastic simulation) may look quantitatively different for θ and for $u(\theta)$, but this is merely induced by the perspective generated by the transformation $u(\theta)$. Let $\pi_\theta(\theta|x)$ denote the posterior probability density of the parameter θ, evaluated at the point θ given the data x. Since finite probability masses are to be preserved under transformation, any (posterior) probability density satisfies $\pi_u(u(\theta)|x) = \pi_\theta(\theta|x)|u'(\theta)^{-1}|$. This means that, in contrast to the situation for likelihood functions, medians and other quantiles of the probability distribution are equivariant, but not the mode or nor the relative parameter location corresponding to a fixed ratio of probability densities. (Expectation values are only approximately equivariant for transformations that are close to their linear approximation over the range where the probability distribution is essentially concentrated.) This means that small sample likelihood inference is essentially different from Bayesian inference, even when diffuse or (almost) flat priors are used.

Both approaches lead commonly to procedures that are *asymptotically* equivalent, in the sense that their relative difference vanishes with increasing sample size. This is related to the fact that asymptotically, together with the influence of the prior distribution, the asymmetry of the likelihood function disappears, and hence also the difference between mean, median, mode, as well as in many cases—due to the central limit theorem and the Gaussian shape of the large-sample likelihood function—the incongruity between quantile-based (Bayesian) and likelihood-ratio based (non-Bayesian) interval estimates.

Furthermore, following the Bayesian paradigm [24, 44, 53, 100, 139] for a moment, it is evident that the probabilistic transformation property should be inherent to the prior probability, since it is not provided by the likelihood function. This can be realised either by imposing 'coherence' on the preferences of any 'rational' man in choosing his (subjective) prior, or by postulating as a working rule a class of priors which automatically satisfies probabilistic equivariance. The latter approach was adopted by Jeffreys [66] who propounded to take as (retrospective) prior the square root of the expected Fisher information $I(\theta) = -E_\theta(\partial^2/\partial\theta^2)l_\theta(\theta; x)$ as a function of the unknown parameter θ. In one-dimensional situations this leads often to quite reasonable, if not completely compelling, results, such as a flat prior for a location parameter or for the logarithm of the scale parameter σ of a normal distribution. (The latter implies a so-called scale-invariant prior, $\pi_J(\sigma^p) \sim (\sigma^p)^{-1}$, for σ to any power p.) On the other hand, for a Poisson distribution $P(\lambda)$ we have $\pi_J(\lambda) \sim \lambda^{-1/2}$ which means a constant prior $\pi_J(\sigma) \sim 1$ for the scale parameter $\sigma = \lambda^{1/2}$, while for the Cauchy ('Lorentz') distribution, where the scale parameter σ is proportional to the full-width at half-maximum of the probability density, we have $\pi_J(\sigma) \sim \sigma^{-1/2}$.

Let us dwell a moment on estimating the probability θ of 'success' on the basis of n Bernoulli experiments, $X_1, \ldots, X_n \sim B(1, \theta)$, to be used in inference on the number of successes in m further experiments, a problem at least 'as ancient as Bayes' [103], and, in a slightly more general setting, 'as recent as Walley' [145]. Jeffreys's rule entails in this situation the prior $\pi_J(\theta) \sim (\theta(1-\theta))^{-1/2}$, which means a flat prior for $u(\theta) = \arcsin(\sqrt{\theta})$. Although reportedly Jeffreys did 'not like' this prior

himself [118], it holds the geometric middle between Laplace's prior for an 'equal distribution of ignorance', $\pi_L(\theta) \sim 1$, and Haldane's prior [58] $\pi_H(\theta) \sim (\theta(1-\theta))^{-1}$, which entails a flat prior for the log odds-ratio $v(\theta) = \log(\theta/(1-\theta))$. All these different priors are of course special cases of the Beta distribution, $Be(\alpha, \beta)$, with density proportional to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for positive α and β, and conjugate to the binomial. (Restricted to integer values of α and β, $Be(\alpha, \beta)$ describes the distribution of the α^{th} order statistic from a sample of size $\alpha+\beta$, corresponding to the physical situation considered by Bayes [16].) Given a choice for α and β, the 'usual' Bayesian rule (minimising the expected posterior quadratic error) is to estimate θ by the posterior expectation $E(\theta|x) = (x+\alpha)/(n+\alpha+\beta)$.

Somewhat in the vein of Fisher, Groningen statisticians are not satisfied with such point estimates. They are also not completely satisfied with the $Be(x+\alpha, n-x+\beta)$ posterior distributions and have developed a theory for distributional inference where attention is restricted to 'weakly unbiased' procedures and, in addition, the integrated risk (with respect to a proper loss function) is minimised. Interestingly, the approach leads to the credence distribution $1/2[Be(x, n-x+1)+Be(x+1, n-x)]$ for the true value of θ, which is close to the $Be(x+1/2, n-x+1/2)$ posterior distribution based on Jeffreys's prior. The weak unbiasedness restriction entails that the distributional decision rule is not admissible with respect to the proper loss function described in [85], and cannot be obtained from a prior distribution by applying Bayes's theorem, see [45], Ch.2. A comparative analysis between various rules of distributional inference for this problem, called 'the fundamental problem of practical statistics' in [103], can be found in [85, 118]. In the following, we restrict our attention again to the less complicated area of point estimation.

Within the Bayesian framework, the posterior median estimator $med(\theta|x)$ (minimising the expected posterior absolute error) has a vantage over $E(\theta|x)$ in that it is equivariant under all monotonic transformations, and (hence) more robust than $E(\theta|x)$ against strong asymmetries. Unfortunately, a closed-form expression for the median of a Beta distribution does not exist. However, a practically reasonable approximation, derived on occasion of writing this paper, is $med(\theta|x) = (x+\alpha-0.3)/(n+\alpha+\beta-0.6)$. The maximum absolute deviation, on log odds scale, between the true and the approximated median is less than 0.01, 0.03, 0.10, 0.15 if both parameters $x+\alpha$ and $n-x+\beta$ are inside the intervals $(1, 9)$, $(0.75, 1000)$, $(0.56, 1000)$, $(0.505, 1000)$, respectively. The expression has been obtained by minimising numerically the sum of squared deviations over the family $(x+\alpha+c)/(n+\alpha+\beta+d)$ as a function of c and d. Somewhat unexpectedly, for $n = 2$ and Laplace's prior ($\alpha = 1$, $\beta = 1$), the posterior median estimator for θ coincides, to a high degree of approximation, with the minimax risk estimator $(x+0.5\sqrt{n})/(n+\sqrt{n})$ for quadratic loss, to be payed 'to the devil', see Steinhaus [136], which makes the estimator an interesting candidate in the small-sample contest [41], with good properties if some (vague) prior information exist that θ is not extreme, say $1/6 < \theta < 5/6$. This digression to 'the fundamental problem of practical statistics' [103], with ramifications to meta-analysis [96], logistic regression [33, 35], contingency table analysis [149], discriminant analysis [63], multivariate calibration [132], and other areas of statistics, has been made to indicate that 'das statistische Spiel [136] noch kein Ende genommen hat', in spite of the convenient framework of the Bayesian paradigm, and despite its aspect of 'a tempest in a teapot' [56]. In a sense, the character of the game is 'open' [136], suggesting a possible useful exploitation, in repetitive situations, of strategies needed by those confronted with the prisoner's dilemma [9].

In more than one dimension, the determinant of Fisher's information matrix has to be taken, which leads in a number of situations to counter-intuitive results for Jeffreys's prior, as was noted by himself, see e.g. [115].

Occasionally, a type of invariance argument to select a specific prior distribution is invoked based on deceptively plausible reasoning. At some instance, in an oral discussion, the first author was confronted with the proposition to use the invariant prior $\pi(\theta) = c/\theta$ for the distribution of any positive, dimensional physical parameter, such as for instance length, based on the grounds that this is the only

density invariant under the choice of physical units (e.g. cm or m). A reference was made to [19]. This argumentation is not compelling, since it is more natural to consider, for each choice of units, coherently, a single member of, for instance, a one-parametric scale family of prior distributions. (The transformation property of probability densities requires that $\pi_{c\theta}(c\theta) = c^{-1}\pi_\theta(\theta)$, but we do not perceive a deductively or practically valid reason why, for an arbitrary parameter restricted only by the requirement to be positive, the function $\pi_{c\theta}$, which is the prior distribution of $c\theta$, should be equal to π_θ.)

In [103], Karl Pearson not only emphasized the practical-theoretical value of but also expressed his concern about Bayes's problem. He spoke about the virtue of "realising that if the views of some authors be correct our superstructure is built, if not on quicksand, at least in uncomfortable nearness to an abyss". On choosing a prior distribution he wrote somewhat jocundly about "taking a bull at his horns" (the peaks of a Beta prior if parameters are less than 1), before essentially deriving the Beta-binomial distribution for r successes in m further experiments at an arbitrary $Be(\alpha, \beta)$ prior distribution, and comparing it with the hypergeometric distribution using the maximum likelihood estimate from the first set of n trials. About the ubiquitous application of the normal approximation, he complains that its "sacrosanct character, in the eyes of many physicists and astronomers, is of the character of a dogma, being based on authority rather than reason". Pearson's analysis of Bayes's problem remains essentially valid, despite considerable advances in computational facilities, and copious more recent investigations, for which we suggest [118, 80, 2] as a possible starting point. As we have seen above, the choice of parameterisation for which a flat prior is postulated remains a serious point of concern, and, moreover, even for flat priors small-sample likelihood inference is essentially different from small-sample Bayesian inference. In view of these considerations, our, somewhat shadowy formulated, extant answer to the first question is: "perhaps, if due attention is paid to the modifying adjective 'approximate' ".

14 What is the Role in Theory and in Practice of Upper and Lower Probabilities?

The expression 'upper' and 'lower' probabilities can be interpreted in various ways, since upper and lower bounds of physical, or of epistemic, probabilities occur in several contexts.

In Groningen a tradition exists, since the mid-seventies, to construct confidence intervals for posterior probabilities [126, 4]. In logistic regression, and even for proportional hazard regression models, similar results can be obtained, see [70, 17]. The standard errors or confidence limits provide an impression of the statistical uncertainties involved in the assessment. Apart from these statistical uncertainties there are the uncertainties caused by systematic errors: parametric models are always wrong to a certain extent. Hence it is somewhat misleading to focus on these statistical uncertainties. Nevertheless they may provide some help in diagnosing the problem addressed to in cryptic issue 1, because overconditioning manifests itself in upper and lower bounds which are very far apart.

In nuclear fusion, a theme of practical and scientific interest is to obtain a reliable estimate of the probability to achieve a burning thermal plasma for which the (internal) alpha-particle heating exceeds the externally supplied heating. Creating such a plasma is of interest for fundamental physical investigation as well as for developing fusion reactor technology. This probability depends on a so-called reference scenario for plasma operation, which is a function of the plasma density, the auxiliary heating power, and some other plasma characteristics such as current, magnetic field and plasma shape, each of which can be controlled within a certain operating domain. In [79], an interval estimate [76] for the confinement time of the ITER tokamak [77], updated for ITER FEAT [10], was transformed, using auxiliary physical information, into a probability estimate for achieving a ratio between internal heating (by helium particles) and external heating larger than a certain lower bound, with special interest to the values 1 and 2.

Further explanation by the last author revealed, however, that a connection with Shafer's theory of

belief measures [130, 144] has been envisaged by the question. Henceforth, we concentrate on this aspect.

Upper and lower bounds of probability measures have been investigated by Dempster [42] and were later interpreted by Shafer [130], in the context of *possibility theory*, as plausibility and belief measures, see also [146, 84, 57]. According to [84], Ch. 7, there is a connection between the membership functions of fuzzy sets and the assignment of possibility and necessity measures, which are special cases of plausibility and belief measures. These are in fact interesting generalisations of probability measures in that the property of additivity is replaced by subadditivity and superadditivity, respectively. A 'strength-of-belief' measure μ_{be} satisfies for instance:

$$\mu_{be}(A \cup B) \geq \mu_{be}(A) + \mu_{be}(B) - \mu_{be}(A \cap B) \tag{1}$$

which implies $\mu_{be}(A) + \mu_{be}(\complement A) \leq 1$. If there is little evidence for A, nor for its complement $\complement A$, one can assign a (very) low degree of belief to both. According to standard Bayesian probability theory, anyone using such (non-additive) degrees of belief would not behave coherently. We have little sympathy with this attitude. Shafer [130] gives an illustrative example why the concept of ignorance is better represented by these super-additive belief measures than in standard Bayesian theory. Consider the following two exhaustive sets of mutually exclusive propositions, which physically presuppose that a planetary orbit is necessary but not sufficient for the existence of life: p_1: There is life in an orbit near Sirius, p_2: There is no life in an orbit near Sirius and q_1: There is life near Sirius, q_2: there are planets, but there is no life near Sirius, q_3: there are not even planets near Sirius. Complete ignorance would be represented by assigning belief measure zero to all these cases. In Bayesian theory one can do little else than assign probability $1/2$ to p_1 and p_2 and probability $1/3$ to q_1, q_2, and q_3. These two assignments, however, are inconsistent since the conjunction of the propositions q_2 and q_3 is equivalent to the proposition p_2. Shafer mentions that such types of inadequacy of representing ignorance has been a reason of the decline of Bayesian probability theory in the nineteenth century [23], see also Zabell [153] for a balance up to 1920.

Because of the additivity property of probability measures, Bayesian inference has a flavour of 'how to gamble if you must' [122], at times an unattractive option to be faced with. Consider a bag with marbles [145]. One is told that each such marble assumes one out of some k colours. Instead of assigning a prior $1/k$ to each of these, one can make a different category, $k+1$, corresponding to 'undefined colour', combine categories, etc., leading to different priors for logically equivalent situations. As argued in [145], representation of prior uncertainty by lower and upper probabilities, satisfies invariance with regard to the sample-space embedding. In the specific setting of multinomial models with Dirichlet priors, they lead to intervals for the unknown parameter that shrink as c/N for some constant c. It is noticed that for $c \propto \sqrt{N}$ the interval shrinks as $1/\sqrt{N}$ and the centre of the interval corresponds to the minimax risk estimator considered by Steinhaus [136], see also [64, 22, 121]. On the other hand, in [145] it is stated that if c depends on N, the principle of coherence is violated. It is questionable, however, whether this principle should be very strictly adhered to, see cryptic issue 10. (Similarly, upper and lower probabilities for the variance or a quantile of the distribution function of the estimator can be calculated.) One interpretation is that they correspond to suprema and infima over classes of prior distributions. From a fundamental point of view, they challenge, however, the Bayesian paradigm. From the theoretical side, in our view, possibility theory has formed an interesting and flexible framework of axiomatised concepts with precise, suggestive interpretations and with analogies from modal (and temporal) logic, see [84, 65, 52], even if we agree with [110] that a monolithic structure for statistical inference does not exist. A simple example is given by the following. In modal logic, the expression $\Diamond p = \sim \Box \sim p$ means that a proposition being possible is equivalent to the statement that its negation is not necessarily true. This corresponds to the property $\mu_{po}(A) = 1 - \mu_{ne}(\complement A)$, meaning that the possibility measure of event A is one minus the necessity measure of the event that A does not occur. (By extension, in possibility theory, the

relation $\mu_{pl}(A) = 1 - \mu_{be}(\complement A)$ holds, with the semantic interpretation that the 'plausibility' of event A to occur is defined as 1 minus our 'belief' that A does not occur.)

To describe its role in practice is more difficult. Fuzzy set theory has led to multifarious applications in various fields, such as control theory, decision theory, pattern recognition, etc., see [84], but a link between theory and empirical practice as strong as between probability theory and statistics for additive measures, is not (yet) so well developed, despite several pioneering efforts, see e.g. [145]. We give three practical examples here.

(I) In analogy with betting behaviour to elicit prior probabilities (see cryptic issue 10), it has been suggested to identify a lower probability with a buying price and an upper probability with a selling price of the same (precious, non-consumer) object by the same person. Consider a Japanese book (hand-)written in Kanji from before the Meiji period, which one is willing to buy for 10.000 ¥ (or less) and to sell for 30.000 ¥ (or more). From this example, it should be apparent that upper and lower probabilities can be quite far apart, and, conceptually more important, they do not correspond at all to upper and lower estimates of the same physical quantity ensuing from incompleteness of the available data.

(II) Upper and lower probabilities can be used as upper and lower limits in sample surveys with non-responder fractions. Suppose that, at some occasion, 40% of the responders voted for proposal A and 60% for proposal B, where the proposals A and B are mutually exclusive, for example a (future) presidential election in the European Union. However, 20% of the population did not vote. In a Bayesian approach, assignment of some prior probability for the non-responder fractions is unavoidable. This can lead to interesting, if also somewhat frictional, discussions, which in our view are not well adapted toward the real issue, since, without further empirical auxiliary information, one simply does not know the opinion of the non-responders. On the other hand, to employ a physical metaphor, in possibility theory, the spectral line 'probability' is split into a doublet (a pair of dual measures). The two lines of the doublet are given different semantic interpretations according to the requirements of the practical situation. In our specific example, we have $\underline{P}(A) = 32\%$, $\underline{P}(\complement A) = 48\%$ as lower probabilities and $\overline{P}(A) = 52\%$, $\overline{P}(\complement A) = 68\%$ as upper probabilities. This is a good starting point for further investigation by sub-sampling or imputation [128] in situations where this is deemed necessary.

(III) Consider as a Gedankenexperiment, albeit not entirely an academic one, the planning of two future plasma fusion devices: (A) a large scale tokamak, designed for testing integrated fusion reactor technology as well as studying the physics of burning plasmas, which are predominantly heated by alpha (i.e. helium) – particles; this device has passed the engineering design phase (six years), and the modalities are being discussed for actual international construction (lasting some ten years), (B) a small scale (high magnetic field) device, essentially designed for studying predominantly alpha-particle heated plasmas, which has passed the conceptual design phase (corresponding to approximately three years), and the question is discussed whether resources should be allocated for a detailed engineering design study. Actual investment costs are high in both cases, but considerably higher in case (A) than in case (B). Here it seems to be justified, concentrating on just the design goal of producing alpha-particle heated plasmas, to estimate the lower probability for machine (A) for achieving its intended fusion performance, and on the upper probability for machine (B), in view of the difference in investment if the objectives are not met. Strict adherence to Bayesian probability theory tends to blur rather than to clarify the actual issue in this situation.

Sub- and super-additive measures are more general and more difficult to handle than additive ones. For instance, the concept of a 'cumulative distribution function' for a one-dimensional, continuous parameter has to be generalised, such that it can be efficiently estimated from empirical data. Upper and lower probabilities bear a minimax character. Conceivably, a semiparametric approach, using for instance Sugeno-type measures [138], which are (sub, super) additive for a single parameter being

(negative, positive) zero, may prove to be worthwhile in describing classes of empirical problems where additive measures are too restrictive, and at the same time allow for effective statistical estimation and procedure evaluation in repetitive situations. Example (II) entails that the quest for additional empirical information, in the vein of Stein's two-step procedure in sequential analysis, may be useful, if feasible, for substantial non-responder fractions in survey sampling [12], beyond the mere assessment of lower and upper probabilities. The true developments in this research area will only become clear in the future. We venture to say that possibly they may emerge from investigations that are not pressurised too much by near-term objectives.

Acknowledgements

The first author gratefully acknowledges correspondence with T.A.B. Snijders from the Sociology Department at Groningen University, the invitation by V. Dose to hold a seminar at the Centre for Interdisciplinary Plasma Physics at MPI Garching, and some useful comments by V. Mukhovatov from the ITER Central Team at JAERI Naka.

References

[1] Ahlbom, A., Bergqvist, U., Bernhardt, J.H. *et al.* (1998). Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields (up to 300 GHz). *Health Physics*, **74**, 494–522.
[2] Albers, C.J. (2003). *Distributional Inference: The Limits of Reason.* Ph.D. Thesis, Groningen University.
[3] Albers, C.J. & Schaafsma, W. (2001). How to assign probabilities if you must. *Statistica Neerlandica*, **55**, 346–357.
[4] Ambergen, A.W. (1993). *Statistical Uncertainties in Posterior Probabilities.* Ph.D. Thesis, Groningen University. Also: CWI Tract No. 93. Amsterdam: Centrum voor Wiskunde en Informatica.
[5] Anderson, T.W. (1984). Estimating linear statistical relationships (The 1982 Wald Memorial Lecture). *The Annals of Statistics*, **12**, 1–45.
[6] Apostol, L. (1972). Illocutionary forces and the logic of change. *Mind*, **81**, 208–224.
[7] Aristotle (330 BC.) *Ethica Nicomachia*, English translation by H. Rackham, Loeb Classical Library No. 73 (1968), text based on the edition by I. Bekker (Berlin, 1831). London: Heinemann.
[8] Austin, J.L. (1975). *How to Do Things with Words,* 2nd ed. Havard University Press.
[9] Axelrod, R. (1987). The evolution of strategies in the iterated prisoner's dilemma. In *Genetic Algorithms and Simulated Annealing: An Overview*, Ed. L. Davis, pp. 32–41. San Mateo, Calif.: Morgan Kaufmann Publishers.
[10] Aymar, R., Chuyanov, V.A., Huguet, M., Shimomura, Y., ITER Joint Central Team & ITER Home Teams. (2001). Overview of ITER-FEAT—the future international burning plasma experiment. *Nuclear Fusion*, **41**, 1301–1310.
[11] Aymar, R. (1999). ITER: an integrated approach to ignited plasmas. *Philosophical Transactions of the Royal Society of London, Series A*, **357**, 471–487.
[12] Beaumont, J.-F. (2000). An estimation method for non-ignorable nonresponse. *Survey Methodology*, **26**, 131–136.
[13] Barlow, R.E. (1992). Introduction to De Finetti (1937). In *Breakthroughs in Statistics*, Eds. N. Johnson and N.L. Kotz, Vol. **I**, pp. 127–133. Heidelberg: Springer-Verlag.
[14] Barnard, G.A. (1992). Introduction to Pearson (1900). In *Breakthroughs in Statistics*, Eds. N. Johnson and N.L. Kotz, Vol. **II**, pp. 1–10. Heidelberg: Springer-Verlag.
[15] Bartlett, M.S. (1949). Probability in mathematics, logic and science. *Dialectica*, **3**, 104–113.
[16] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London, Series A*, **53**, 370–418. Reprinted in Biometrika (1958) **45**, 293–315 with an introduction by G.A. Barnard.
[17] Beukhof, J.R., Kardaun, O.J.W.F., Schaafsma, W., Poortema, K., Donker, A.J.M., Hoedemaeker, P.J. & van der Hem, G.K. (1985). Toward individual prognosis of IgA nephropathy. *Kidney International*, **29**, 549–556.
[18] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1998). Semiparametrics. In *Encyclopedia of Statistical Sciences*, Eds. S. Kotz, C.B. Read and D.L. Banks, Update Vol. **2**. New York: Wiley.
[19] Berger, J.O. (1980). *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. Heidelberg: Springer-Verlag.
[20] Berger, J. & Wolpert, R. (1988). *The Likelihood Principle,* 2nd ed. Hayward, Calif.: Institute of Mathematical Statistics.
[21] Bernoulli, D. (1778). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction (with comment by L. Euler). *Acta Academiae Petropolitanae*, 3–33. English translation by C.G. Allen and an introductory note by M.G. Kendall in *Biometrika* (1961) **48**, 1–18.
[22] Blackwell, D. & Girshick, M.A. (1954). *Theory of Games and Statistical Decisions.* New York: Wiley. Reprinted (1979). New York: Dover.
[23] Boole, G. (1854). *An Investigation of the Laws of Thought, on Which are founded the Mathematical Theories of Logic and Probabilities.* London: Walton and Maberly. Reprinted (1976). New York: Dover.
[24] Box, E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley.
[25] DeBruin, R., Salomé, D. & Schaafsma, W. (1999). A semi-Bayesian method for nonparametric density estimation.

Computational Statistics and Data Analysis, **30**, 19–30.
[26] Cochran, W.G. (1977). *Sampling Techniques.* 3rd ed. New York: Wiley.
[27] Cordey, J.C., Goldston, R.J. & Parker R.R. (1992). Progress toward a tokamak fusion reactor. *Physics Today*, Jan. 1992, 22–30.
[28] Caroll, R.J., Rupert, D. & Stefanski, L.A. (1995). *Measurement Error in Non-linear Models*. London: Chapman and Hall.
[29] Cox, D.R. (1958). Some problems connected with statistical evidence. *Annals of Mathematical Statistics*, **29**, 357–372.
[30] Cox, D.R. & Miller, H.D. (1965). *The Theory of Stochastic Processes.* London: Methuen.
[31] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, 187–220.
[32] Cox, D.R. (1975). Partial Likelihood. *Biometrika*, **66**, 188–190.
[33] Cox, D.R. (1989). *The Analysis of Binary Data,* 2nd ed. Monographs on Statistics and Applied Probability, **32**. London: Chapman and Hall.
[34] Cox, D.R. (1992). Causality: Some Statistical Aspects. *Journal of the Royal Statistical Society B*, **155**, 291–301.
[35] Cox, D.R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441–461.
[36] Cox, D.R. (1997). *The Nature of Statistical Inference.* Report TW W-9708, Groningen University.
[37] Cox, D.R. (1997). The current position of statistics: a personal view (with discussion). *International Statistical Review*, **65**, 261–290.
[38] Cox, D.R. & Reid, N. (2000). *Theory of the Design of Experiments*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.
[39] Dagum, E.B., Cholette, A. & Chen, Z.-G. (1998). A unified view of signal extraction, benchmarking, interpolation and extrapolation in time series. *International Statistical Review*, **66**, 245–269.
[40] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaŕe*, **7**, 1–68. Reprinted in *Studies in Subjective Probability*, Eds. H.E. Kyburg Jr. and H.E. Smokler (1980) pp. 93–158. New York: Dover.
[41] Dehling, H.G., Dijkstra, T.K., Guichelaar, H.J., Schaafsma, W., Steerneman, A.G.M., Wansbeek, T.J. & Van der Zee, J.T. (1996). Structuring the inferential contest. In *Bayesian Analysis in Statistics and Econometrics*, Eds. D.A. Berry, K.M. Chaloner and J.K. Geweke, Ch. 46, 539–547. New York: Wiley.
[42] Dempster, A.P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, **38**, 325–339.
[43] Doksum, K.A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *The Annals of Statistics*, **15**, 325–345.
[44] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.
[45] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.
[46] Fienberg, S.E. & Hinkley, D.V. (Eds.) (1980). *R.A. Fisher: An Appreciation.* Lecture Notes in Statistics, Vol. 1. Heidelberg: Springer-Verlag.
[47] Fisher, R.A. (1935). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, **98**, 39–82.
[48] Fisher, R.A. (1973). *Statistical Methods and Statistical Inference*, 3rd ed. New York: Hafner. First ed. (1956). Edinburgh: Oliver and Boyd.
[49] Fisher, R.A. (1970). *Statistical Methods for Research Workers*, 14th ed. Edinburgh: Oliver and Boyd.
[50] Fishburn, P. (1999). Preference structures and their numerical representations. *Theoretical Computer Science*, **217**, 359–383.
[51] Friedman, N. & Halpern, J.Y. (1997). Modelling belief in dynamical systems. Part I: Foundations. *Artificial Intelligence*, **95**, 257–316.
[52] Friedman, N. & Halpern, J.Y. (1999). Modelling belief in dynamical systems. Part II: Revision and update. *Journal of Artificial Intelligence Research*, **10**, 117–167.
[53] Garrett, A.J.M. (1989). Probability, philosophy and science: a briefing for Bayesians. In *Maximum Entropy and Bayesian Methods*, Ed. J. Skilling. Dordrecht: Kluwer.
[54] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.* Hamburg: Perthes & Besser. *Werke*, **7**, 1–280. English translation by H.C. Davis as Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections (1963). New York: Dover.
[55] Gauss, C.F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.* Pars prior, postertior et supplementum. Königliche Gesellschaft der Wissenschaften zu Göttingen. *Werke*, **4**, 3–26, 29–53 and **4**, 104–108. English translation by G.W. Stewart as Theory of the Combination of Observations Least Subject to Errors (1995). Portland, OR.: Book News Inc.
[56] Geisser, S. (1984). On prior distributions for binary trials. *American Statistician*, **38**, 244–251.
[57] Guan, J.W. & Bell, D.A. (1991–1992). *Evidence Theory and its Applications.* (Vols 1 and 2). New York: North Holland.
[58] Haldane, J. (1931). A note on inverse probability. *Proceedings of the Cambridge Philosophical Society*, **28**, 55–61.
[59] Hanfelt, J.J. & Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461–477.
[60] Hemelrijk, J. (1960). The statistical work of David van Dantzig (1900–1959). *Annals of Mathematical Statistics*, **31**, 269–275.
[61] Heij, C. & Willems, J.C. (1989). A deterministic approach to approximate modelling. In *From Data to Model*, Ed. J.C. Willems. New York: Springer-Verlag.

[62] Hillegers, L.T.M.E. (1986). *The Estimation of Parameters in Functional Relationship Models.* Ph.D. Thesis, Eindhoven University.
[63] Hirst, D.J., Ford, I. & Critchley, F. (1990). An empirical investigation of methods for interval estimation of the log odds ratio in discriminant analysis. *Biometrika*, **77**, 609–615.
[64] Hodges, J.L., Jr. & Lehmann, E.L. (1950). Some problems in minimax point estimation. *Annals of Mathematical Statistics*, **21**, 182–197.
[65] Hughes, G. & Cresswell, M. (1996). *A New Introduction to Model Logic.* London: Routledge.
[66] Jeffreys, H. (1998). *Theory of Probability,* 3rd ed. Oxford University Press.
[67] Joyce, J.M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, **65**, 575–603.
[68] Kalbfleisch, J.D. (1975). Sufficiency and conditionality (with discussion). *Biometrika*, **62**, 251–268.
[69] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.
[70] Kardaun, O.J.W.F. (1986). *On Statistical Survival Analysis.* Ph.D. Thesis, Groningen University. Chapters 2 and 3 also in *Handbook of Statistics*, Eds. C.R. Rao and R. Chakraborty, (1991) Vol **8**, 407–459 and 461–513. Amsterdam: Elsevier.
[71] Kardaun, O.J.W.F., McCarthy, P.J., Lackner, K., Riedel, K. & Gruber, O. (1987). A statistical approach to plasma profile invariance. In *Theory of Fusion Plasmas*, Eds. A. Bondeson, E. Sindoni and F. Troyon, pp. 435–444. Bologna: EUR 11336 EN.
[72] Kardaun, O.J.W.F. for the H-Mode Database Working Group (1993). ITER: analysis of the H-mode confinement and threshold databases. *Plasma Physics and Controlled Nuclear Fusion Research, Proceedings of the* 14^h *International Conference on Fusion Energy* (*Würzburg 1992*) Vol **3**, 251–270. Vienna: IAEA.
[73] Kardaun, O.J.W.F., Kardaun, J.W.P.F., Itoh, S.-I. & Itoh, K. (1992). Discriminant analysis of plasma fusion data. In *Computational Statistics X*, Eds. Y. Dodge and J. Whittaker, Vol **1**, 163–170. Heidelberg: Physica-Verlag.
[74] Kardaun, O.J.W.F., Kus A. *et al.* (1996). Generalising regression and discriminant analysis: catastrophe models for plasma confinement and threshold data. In *Computational Statistics XII*, Ed. A. Prat, Vol **1**, 313–318. Heidelberg: Physica-Verlag.
[75] Kardaun, O.J.W.F. (1997). Some answers to the 'cryptic issues for discussion' posed by Professor D.R. Cox on occasion of his Bernoulli Lecture at Groningen University. *IPP Internet Report* 97/5 1.2. Available on internet `http://www.ipp.mpg.de/netreports`. Garching: Max-Planck-Institut für Plasmaphysik.
[76] Kardaun, O.J.W.F. (1999). Interval estimation of global H-mode energy confinement in ITER. *Plasma Physics and Controlled Fusion*, **41**, 429–469.
[77] ITER Physics Basis Document (1999). *Nuclear Fusion*, **39**, 2173–2664.
[78] Kardaun, O.J.W.F. for the International Confinement Database Working Group (2000). Next-step tokamak physics: confinement-oriented global database analysis. *Proceedings of the 18th IAEA Conference on Fusion Energy (Sorrento 2000)*, IAEA-CN-77-ITERP/04, available on CD-ROM (ISSN 1562-4153) and on internet `http://www.iaea.org/programmes/ripc/physics/fec2000/html/node238.htm`. Vienna: IAEA.
[79] Kardaun, O.J.W.F. (2002). On estimating the epistemic probability of realizing $Q = P_{fus}/P_{aux}$ larger than a specified lower bound in ITER. *Nuclear Fusion*, **42**, 841–852.
[80] Kardaun, O.J.W.F. & Schaafsma, W. (2003). *Distributional Inference: Towards a Bayes–Fisher–Neyman Compromise.* Available on request.
[81] Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.
[82] Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**, 887–906.
[83] Klein, J.P & Moeschberger, M.L. (2003). *Survival Analysis,* 2nd ed. New York: Springer.
[84] Klir, G.J. & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* New Jersey: Prentice Hall.
[85] Kroese, A.H., Van der Meulen, E.A., Poortema, K. & Schaafsma, W. (1995). Distributional inference. *Statistica Neerlandica*, **49**, 63–82.
[86] Kyburg, H.E., Jr., (1983). The reference class. *Philosophy of Science*, **50**, 374–397.
[87] Leonov, V.M. & Chudnovskij, A.N. (2003). Dependence of the energy confinement in the L- and H-modes on the tokamak aspect ratio. *Plasma Physics Reports*, **29**, 97–104.
[88] Li, B. & McCullagh, P. (1994). Potential functions and conservative estimating functions. *Annals of Statistics*, **22**, 340–356.
[89] Lindley, D.V. (1985). *Making Decisions,* 2nd ed. New York: Wiley.
[90] Linssen, H.N. (1980). *Functional Relationships and Minimum Sum Estimation.* Ph.D. Thesis, Eindhoven University.
[91] Ljung, L. (1987). *System Identification, Theory for the User.* New Jersey: Prentice Hall.
[92] Loève, M. (1963). *Probability Theory,* 3rd ed. New Jersey: Van Nostrand.
[93] Mannoury, G. (1934). Die signifischen Grundlagen der Mathematik. *Erkenntnis*, **4**, 288–309.
[94] McCarthy, P.J., Riedel, K., Kardaun, O.J.W.F., Murmann, H.D., Lackner, K. & the ASDEX Team (1991). Scalings and plasma profile parameterisation of ASDEX high-density Ohmic discharges. *Nuclear Fusion*, **31**, 1595–1633.
[95] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed. Monographs on Statistics and Applied Probability, **37**. London: Chapman and Hall.
[96] Mosteller, F. & Chalmers T.C. (1992). Meta-Analysis: methods for combining independent studies. Some progress and problems in meta-analysis of clinical trials. *Statistical Science*, **7**, 227–236.
[97] Mukhovatov, V. *et al.* (2003). Comparison of ITER performance prediction by semi-empirical and (partly) theory-based transport models. *Proceedings of the 19th IAEA Conference on Fusion Energy (Lyon, 2002)*, IAEA-CN-94-CT/P03, submitted for publication.
[98] Neyman, J. (1961). Silver jubilee of my dispute with Fisher. *Journal of the Operations Research Society of Japan*, **3**, 145–154.

[99] Neyman, J. (1967). R.A. Fisher (1980–1962): An appreciation. *Science*, **156**, 1456–1460.
[100] O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Vol 2B* (*Bayesian Inference*). Oxford University Press.
[101] Ongena, J. & Van Oost, G. (2000). Energy for future centuries: will fusion be an inexhaustible, safe and clean energy source? *Fusion Technology*, **37**, 3–15.
[102] Parcival D.B. & Walden, A.T. (1993). *Spectral Analysis for Physical Applications.* Cambridge University Press.
[103] Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika*, **13**, 1–16.
[104] Pearson, E.S. & Neyman, J. (1933). On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, **231**, 289–337.
[105] Polderman, J.W. & Willems, J.C. (1998). *Introduction to Mathematical Systems Theory: A Behavioral Approach.* Heidelberg: Springer-Verlag.
[106] Preuss, R., Dose, V. & Von der Linden, W. (1999). Dimensionally exact form-free energy confinement scaling in W7-AS. *Nuclear Fusion*, **39**, 849–862.
[107] Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**, 300–325.
[108] Rao, C.R. (1973). *Linear Statistical Inference and its Applications.* Second edition. Wiley: New York. (Paperback edition: 2001).
[109] Rao, C.R. (1990). Personal communication to W. Schaafsma.
[110] Rao, C.R. (1992). R.A. Fisher: The founder of modern statistics. *Statistical Science*, **7**, 34–48.
[111] Rao, C.R., Schaafsma, W., Steerneman, A.G.M. & Van Vark, G.N. (1993). Inference about the performance of Fisher's linear discriminant function. *Sankhyã, Series B*, **55**, Pt 1, 27–39.
[112] Reichenbach, H. (1916). *Der Begriff der Wahrscheinlichkeit für die mathematische Darstellung der Wirklichkeit.* Ph.D. Thesis, Erlangen University. Also in: Zeitschrift für Philosophie und philosophische Kritik, Bd 161–163.
[113] Reichenbach, H. (1935). *Wahrscheinlichkeitslehre: eine Untersuchungüber die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung.* English translation: (1948). *The Theory of Probability, an Enquiry into the Logical and Mathematical Foundations of the Calculus of Probability.* Berkeley, Calif.: University of California Press.
[114] Reid, N. (1995). The roles of conditioning in inference. *Statistical Science*, **10**, 138–157.
[115] Robert, C.P. (1994). *The Bayesian Choice.* Heidelberg: Springer Verlag.
[116] Ruben, H. (2002). A simple conservative and robust solution of the Behrens-Fisher problem. *Sankhyã A*, **64**, 139–155.
[117] Ryter, F. for the H-mode Database Working Group. (1996). H-mode power threshold database for ITER. *Nuclear Fusion*, **36**, 1217–1264.
[118] Salomé, D. (1998). *Statistical Inference via Fiducial Methods.* Ph.D. Thesis, Groningen University.
[119] Särndal, C.-E., Swensson, B. & Wretman, J. (1997). *Model Assisted Survey Sampling.* New York: Springer-Verlag.
[120] SAS Institute Inc. (1993). *SAS/ETS User's Guide*, Version 6, 2nd ed. North Carolina, Cary.
[121] Savage, L.J. (1972). *The Foundations of Statistics.* 2nd ed. New York: Dover. First ed. (1954). New York: Wiley.
[122] Savage, L.J. (1965). *How to Gamble if You Must: Inequalities for Stochastic Processes.* New York: McGraw-Hill.
[123] Savage L.J. (1971). Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association*, **66**, 783–801.
[124] Savage, L.J. (1976). On rereading R. A. Fisher (with discussion). *The Annals of Statistics*, **4**, 441–500.
[125] Schaafsma, W. & Smid, L.J. (1966). Most stringent somewhere most powerful tests against alternatives restricted to a number of linear inequalities. *Annals of Mathematical Statistics*, **37**, 1161–1172.
[126] Schaafsma, W. (1985). Standard errors of posterior probabilities and how to use them. In *Multivariate Analysis*, Ed. P.R. Krishnaiah, Vol. **6**, 527–548.
[127] Schaft, A. (1984). *System Theoretic Description of Physical Systems.* Ph.D. Thesis, Groningen University. Also: CWI Tract No. 3. Amsterdam: Centrum voor Wiskunde en Informatica.
[128] Schulte Nordholt, E. (1998). Imputation: methods, simulation experiments and practical examples. *International Statistical Review*, **66**, 157–180.
[129] Schwartz, D., Flamant, R. & Lellouch, J. (1980). *Clinical Trials*. London: Academic Press.
[130] Shafer, G. (1976). *A mathematical theory of evidence.* New Jersey: Prentice University Press.
[131] Shimomura, Y., Murakami, Y., Polevoi, A.R., Barabaschi, P., Mukhovatov, V. & Shimada, M. (2001). ITER: opportunity of burning plasma studies. *Plasma Physics and Controlled Fusion*, **43**, A385–A394.
[132] Smilde, A.K., Van der Graaf, P.H., Doornbos, D.A., Steerneman, T. & Sleurink A. (1990). Multivariate calibration of reversed-phase chromatographic systems—Some designs based on 3-way data-analysis. *Analytica Chimica Acta*, **235**, 41–51.
[133] Snipes, J. for the International H-mode Threshold Database Working Group (2000). Latest results on the H-mode threshold database. *Plasma Physics and Controlled Fusion*, **42**, A299–A308.
[134] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society, Series A*, **157**, 357–416.
[135] Stein, C. (1981). Estimation of the mean of a multivariate distribution. *The Annals of Statistics*, **9**, 1135–1151.
[136] Steinhaus, H. (1957). The problem of estimation. *Annals of Mathematical Statistics*, **28**, 633–648.
[137] Stigler, S.M. (1999). Gauss and the invention of least squares. In *Statistics on the Table: The History of Statistical Concepts and Methods.* Cambridge MA: Havard University Press.
[138] Sugeno, M. (1977). Fuzzy measures and fuzzy integrals: A survey. In *Fuzzy Automata and Decision Processes*, Eds. M.M. Gupta, G.N. Saridis and B.R. Gaines. New York: North-Holland.
[139] Van Dantzig, D. (1957a). Statistical priesthood (I) (Savage on personal probabilities). *Statistica Neerlandica*, **11**, 1–16.
[140] Van Dantzig, D. (1957b). Statistical priesthood (II) (Sir Ronald on scientific inference). *Statistica Neerlandica*, **11**, 185–200.
[141] Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

[142] Van der Waerden, B.L. & Smid, L.J. (1935). Eine Axiomatik der Kreisgeometrie. *Mathematische Annalen*, **110**, 753–776.
[143] Wald, A. (1950). *Statistical Decision Functions.* New York: Wiley.
[144] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Monographs on Statistics and Applied Probability, **42**. London: Chapman and Hall.
[145] Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 3–57.
[146] Walley, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence*, **83**, 1–58.
[147] Weigend, A.S. & Gerschenfeld, N.A. (editors) (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past.* Santa Fe Institute Studies in the Sciences of Complexity. New York: Addison-Wesley.
[148] Welby, V. (1896). Sense, Meaning and Interpretation. Parts I & II. *Mind*, **5**, 24–37 & 186–202.
[149] Wermuth, N. & Cox, D.R. (1998). On the application of conditional independence to ordinal data. *International Statistical Review*, **66**, 181–199.
[150] Whitehead, A.N. (1967). *Science and the Modern World.* New York: Free Press.
[151] Wong, W.H. (1986). Theory of partial likelihood. *The Annals of Statistics*, **14**, 88–123.
[152] Yushmanov, P.N., Takizuka, T., Riedel, K.S., Kardaun, O.J.W.F., Cordey, J.G., Kaye, S.M. & Post, D.E. (1990). Scalings for tokamak energy confinement. *Nuclear Fusion*, **30**, 1999–2006.
[153] Zabell, S.L. (1989). R.A. Fisher on the history of inverse probability. *Statistical Science*, **4**, 247–256.
[154] Zabell, S.L. (1992). R.A. Fisher and the fiducial argument. *Statistical Science*, **7**, 369–387.
[155] Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, **67**, 45–69.

Résumé

Cette étude contient la formulation originaire et des réponses conjointes par un groupe de scientifiques formés en statistique sur quattorze 'problèmes cryptiques pour discuter' posés au public par Professor Dr. D.R. Cox après son Discours de Bernoulli en 1997 à l'Université de Groningue.

Discussion

Tore Schweder

University of Oslo, Norway
E-mail: tore.schweder@econ.uio.no

Statistics is marred, or perhaps, blessed with mysterious and difficult questions. Such cryptic issues must be discussed to keep statistics a vital methodological discipline. Some of these issues will find their resolution, while others will be with us for a long time. The latter might be regarded as metaphysical or eternal questions, and some may find it stupid to waste time on questions that cannot be resolved. However, some of these questions are so central to the understanding of statistical work that they cannot be avoided. They are simply too nagging to the thinking student and scientist to simply be fended off. We have to keep discussing them to keep fit, and to be of interest. Such discussions, as that in the paper, are also quite enjoyable. By formulating his cryptic issues, Cox has served us 14 balls that the Groningen statisticians have put into play. I basically agree to the views that have been put forward in the paper. My comments are mainly additional.

1. How is overconditioning to be avoided? Conditioning is a central part of statistical inference. Models should be conditioned on prior information, and without statistical models there will be no statistical inference. In statements of uncertainty, Fisher, (1973, p. 58) found it "essential to take the whole of the data into account." He includes prior information in "the whole of the data". Relevant knowledge of the Dutch balls in the urn considered by the authors is prior data along with the outcome of the sampling experiment, and it must all be taken into account. If this knowledge guides us to a "recognisable subset" (Fisher 1973) or "reference set" [92] of urn sampling schemes for which a frequency based probability model can be established, frequentist uncertainty statements like confidence intervals are within reach. If, on the other hand, the relevant knowledge makes the experiment

unique in the sense that repetition is hard to envisage, as seems to be the case for the Dutch ball sampling experiment, frequentist methods are of no help, although the observed data help to reduce the uncertainty (at least 25 of the Dutch balls must be black). Personalistic probability modelling and inference might however be done. The question is then how, and to whom, the inference is usable or meaningful.

Neglecting relevant aspects of the data or the prior information does not solve the problem of overconditioning, at least not in the scientific context. Here, inference must be based on the data, the whole of the data, and nothing but the data. The provision that the data are relevant for the question at hand is admittedly difficult, and must be subject to judgement. If conditioning, in the sense of taking prior information into account, precludes acceptable frequency based modelling, the frequentist statistician must give up. Personalistic inference might still be of scientific interest. Inference based on models is at least internally consistent. Its scientific impact when based on a personalistic model and/or prior distributions on parameters really depends on the status of the person or the group of persons that make the judgements behind the model and the priors. The decision context is different from the scientific context, and certainly more open for personalistic inference and selective use of data.

2. How convincing is "the strong likelihood principle"? Not much if understood strictly, and mostly for the reasons given by the authors. The study protocol or the experimental design might influence the inference. If, say, the size of the data depends on the parameter, this aspect should be accounted for in the confidence intervals. Consider a mark-recapture experiment, where captures are made according to a Poisson process with intensity proportional to population size, which is to be estimated. Several protocols could be imagined. Schweder (2003) found that the confidence distribution (Schweder & Hjort, 2002) is less peaked and less skewed when the protocol allows the numbers of captures to grow (stochastically) with the population size, say when the experiment has a fixed budget, than when the numbers of captures are fixed in advance. To disregard that data tend to be more plentiful under a fixed budget protocol the higher the abundance is, say of whales, contradicts that our confidence interval shall cover the parameter (number of whales) with given probability regardless of its size. The study protocol should specify a strategy for how data are gathered. Our model should be conditioned on the strategy, and thus also reflect the behaviour of the observer. When this can be fully expressed in the statistical model, the likelihood might hold all the information. But often, this is not possible, and additional aspects of the study protocol must be accounted for in the analysis.

Another side of this issue is that the model, and thus the likelihood function, might depend on the question at hand, even when the prior information behind the model is given. Hjort & Claeskens (2003) propose a 'focused information criterion' for model selection when a particular parameter is of primary importance. This criterion leads to different models than do popular omnibus criteria such as Akaike's.

3. What is the role of probability in formulating models for systems, such as economic time series, where even hypothetical repetition is hard to envisage? The authors make a good case for statistical system identification algorithms, stochastic modelling, and simulation as pragmatic ways of formulating simple approximate models for complex processes, and to assess lack of fit between model and observed data.

Stochastic computer simulation was not an option to T. Haavelmo when he started the probability revolution in econometrics. He found the language and theory of probability to be well suited for the description of observations, for formulating theory, for expressing uncertainty, and for statistical analysis in economics. "What we want are theories that, without involving us in direct contradictions, state that the observations will *as a rule* cluster in a limited subset of the set of all conceivable observations, while it is still consistent with the theory that an observation falls outside this subset

"now and then". As far as I know, the scheme of probability and random variables is, at least for the time being, the only scheme suitable for such theories." (Haavelmo, 1944, p. 40).

Probability continues to be the best scheme. Haavelmo insisted on consistency in theory. He would presumably also appreciate consistency in methodology. The validation of methods by simulation testing might be regarded as a check on consistency in methodology. Haavelmo was critical to the various types of foundations of probability. To him, rigorous notions of probabilities exist "only in our rational minds, serving us only as a tool for deriving practical statements". However, he appears to be leaning towards a personalistic interpretation of probability for "economic time series where a repetition of the "experiment" is not possible or feasible. Here, we might then, alternatively, interpret "probability" simply as a measure of our *a priori confidence* in the occurrence of a certain event." (Haavelmo, 1944, p. 48).

Haavelmo's paper was path-breaking and influential, and won him the Nobel Prize in economics. In his paper he explained how the Neyman–Pearson theory applies to econometrics. The frequentist view of Neyman and Pearson might have overshadowed Haavelmo's more personalistic view, and the optimality results might have helped to establish the frequentist paradigm in econometrics. Optimal behaviour is certainly a virtue in economics! Econometricians are still mostly frequentists, but the personalistic view of probability, especially for "experiments of Nature" that cannot be repeated, seems to be advancing. Poirier (1988) argues in favour of the Bayesian view in economics, mostly on logical grounds akin to Haavelmo's view. Economics is a 'speculative' science in the sense that the class of admissible probability models in most cases is very large. The real aim might be more to bring some order and insight into the bewildering reality we observe, than to make predictions with controlled or optimal frequentist properties. The aim is of course also to hit well with the predictions. The Bayesian view might thus win out in the long run, and the population of "heathens" that Poirier (1988) refers to might go extinct in economics.

The authors point to the good agreement that has been experienced between the standardised prediction errors and the normal distribution they were supposed to follow for the Dutch economy from 1981 to 1990. They modestly write that the "standard errors were not meaningless". Certainly not. But was that due to the stochasticity in the model having its origin in the real world rather than in the mind of the data analyst? Stochasticity is a mental construct. That the frequency distribution of the prediction errors agrees with the model reflects good judgement on the part of the Dutch econometricians, or good luck, whether God throws dice or has determined the Dutch economy according to a complicated plan. That probability statements are personalistic does not withhold them from trial. There are good, and there are bad judgements. To the extent that future data allow the personalistic probability statements to be evaluated, the good judges can be separated from the poor.

5. Is it fruitful to treat inference and decision analysis somewhat separate? Yes, indeed. At least when inference is understood as making scientific statements, while decision analysis means analysis for the purpose of making decisions. Science is about truth, while decisions are about what to do. In science, the task is to understand and to broaden our common and public knowledge. Bad science represents a loss to the public at large. Decision-making is more of a private undertaking, and bad decision leads to unnecessary losses for the decision maker. Decisions must be taken whether data are good or bad, but in science we are reluctant to make statements when data are bad. Personalistic probability is in place in the decision context where one is forced to act and where the losses are private, while frequentist inference is more fitting in the scientific context where the truth is assumed permanent and active also in repetitions. Science does, of course, give guidance to decision makers, and scientists make decisions. There is a continuum between the two poles of inference and decision analysis, but it is both fruitful and correct to keep the poles apart.

7. Are all sensible probabilities ultimately frequency based? When following the authors in using

words distinct from 'probability' for each type of uncertainty not based on any frequency aspect, personalistic probability statements would be personal claims regarding frequencies in hypothetical repetitions. This seems too restrictive. Personal probability might represent the 'degree of belief' even in cases when repetitions are impossible. Repetitions are often difficult, not the least in the social sphere where the reality is changing more rapidly than in nature, and where it even might respond to the results of a study. Personalistic probability statements might be in place in such situations, but they are not entirely exempt from trial. If they are in bad accordance with observed frequencies in broadly similar situations, they will be judged as misguided. Are these probability statements then ultimately frequency based?

Sidestepping the issue, one might ask whether statements about uncertainty based on frequentist inference should be denoted by something different from probability. Degree of confidence is a well-established term, and we teach our students to keep it apart from probability, even though it certainly is frequency-based. Fisher (1973, p. 59) discusses his concept of fiducial probability. Referring to the origin of the concept in 1930, and to our ordinary frequency concept, he states "For a time this led me to think that there was a difference in logical content between probability statements derived by different methods of reasoning. There are in reality no grounds for any such distinction." Fisher's fiducial probability has not caught on. This is unfortunate, and might partly be due to Fisher's insistence on this concept being identical to the ordinary frequency based probability. Neyman (1941) interpreted fiducial probability as degree of confidence. This leads to the term 'confidence distribution' instead of fiducial probability distribution (Efron, 1998; Schweder & Hjort, 2002). A confidence distribution distributes confidence over interval statements concerning a parameter, and makes no claim concerning frequencies. A confidence distribution can hardly lead to other than interval statements, and should therefore not be understood as representing a sigma-additive distribution. A family of simultaneous confidence regions indexed by degree of confidence, for example confidence ellipsoids based on the normal distribution, leads to a multivariate confidence distribution. Since it only can be interpreted as a representation of confidence region statements, it should not be treated as a sigma-additive measure, and one cannot in general reduce dimensionality in a multivariate confidence distribution by simple marginalisation. Both for this methodological reason, and because of its ontological status, 'confidence' should be kept apart from 'probability'.

11. How useful is a personalistic theory as a base for public discussion? I will comment on this issue with reference to public discussion of economic matters. Since economics can be regarded as a speculative science, the stochastic models aim not so much at the truth as at establishing a consistent basis for analysis and discourse. The use of personalistic priors would then not be more of a hindrance for public economic debate than using stochastic models based on judgements. Without stochastic models, econometrics is impossible, as Haavelmo (1944) put it. Poirier (1988) cites an exchange between leading economists:

Klamer:	"Are you after the truth?"
Lucas:	"Yeh. But I don't know what we mean by truth in our business."
Sargent:	"Listening to you, to Lucas, and the others . . . I cannot help wondering what economic truth is?"
Tobin:	(Laughter) "That is a deep question. As far as macroeconomics is concerned, my objective has been to have models in which behaviour is assumed to be rational, in which the gross facts of economic life are explained, and which may not give great forecasts in a technical sense but at least an understanding of what is happening and what is going to happen."

Public debate with quantitative arguments is certainly widespread and important in economic matters. These debates are more about what to do than what is true. In this sense, public debate in economics has more to do with decision analysis than with scientific inference, and in the context of

decision-making, personalistic views are certainly in place. The frequentist notions of a permanent truth and of forecasts in a "technical sense" are difficult in economics, as the leading economists remark. It is in this context of interest to note that replication of experiments or studies are rare in economics. Backhouse (1997) comments on this, and notes that replication in economics seldom is more than checking the data analysis, without gathering new data. This is at least the picture that emerges from what leading journals publish.

13. Do rather flat priors lead the Bayesian to approximate confidence intervals, or do they have other justifications? Sidestepping the issue, will really reference priors lead to good confidence intervals? They certainly often do in situations without nuisance parameters. In multi-parameter situations, they might however go badly wrong. Consider a simple example with a normally distributed p-dimensional observed vector. The only parameter is the free p-dimensional location parameter μ. With a flat reference prior, the joint posterior is the normal distribution located at the observed vector. Posteriors for derived parameters are then obtained by integration. For non-linear parameters such as $\theta = \|\mu\|^2$, the posterior distribution is biased in the frequentist sense that the coverage probability for a credibility interval is different from the nominal credibility. Is there then something wrong with the flat reference prior? Should the prior be chosen with a view to the parameter of interest, and be different for different parameters? Yes, perhaps. This would then be in accordance with the fact that the multivariate normal confidence distribution cannot in general provide confidence distributions for derived parameters by mere integration. Requiring unbiasedness has its price!

References

Backhouse, R.E. (1997). *Truth and Progress in Economic Knowledge*. Cheltenham: Edward Elgar Publishing Limited.
Efron, B, (1998). R.A. Fisher in the 21st Century (with discussion). *Statistical Science*, **13**, 95–122.
Fisher, R.A. (1973). *Statistical methods and scientific inference*. (Third Edition). Toronto: Hafner Press.
Haavelmo, T. (1944). The probability approach to econometrics. *Econometrica*, **12**, Supplement. 120 pages.
Hjort, N.L. & Claeskens, G. (2003). The AIC and the FIC. *J. Amer. Statist. Assoc.*, in review.
Neyman, J. (1941). Fiducial Argument and the Theory of Confidence Intervals. *Biometrika*, **32**, 128–150.
Poirier, D.J. (1988). Frequentist and Subjectivist Perspectives on the Problem of Model Building in Economics. *Journal of Economic Perspectives*, **2**, 121–144.
Schweder, T. (2003). Abundance estimation from multiple surveys: confidence distributions and reduced likelihoods for bowhead whales off Alaska. *Biometrics*, in review.
Schweder, T. & Hjort, N.L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, **29**, 309–332.

Discussion

José M. Bernardo

Universidad de Valencia, Spain
E-mail: jose.m.bernardo@uv.es

From a traditional, mostly frequentist, viewpoint the authors have formulated some sort of *communis opino* on the 'cryptic' issues posed by Professor Cox. In a somewhat telegraphic style, an alternative approach to those issues is described below from a very different viewpoint. In the first section, some required foundational points are discussed; the second section contains a possible Bayesian approach to each of the 14 issues discussed.

Foundations

1. 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' 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 $L\{\pi_\theta(. \mid \boldsymbol{z}), \boldsymbol{\theta}\} = -\log[\pi(\boldsymbol{\theta} \mid \boldsymbol{z})]$ associated to a posterior density $\pi_\theta(. \mid \boldsymbol{z})$.

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

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

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 $\boldsymbol{t}^*(\boldsymbol{z}) = \lim_{n \to \infty} \boldsymbol{t}_n^b(\boldsymbol{z})$.

2. 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),

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

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 $R_t(\boldsymbol{\theta} \mid L)$ 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 a 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 $\forall \boldsymbol{\theta} \in \Theta$, $R_{t_1}(\boldsymbol{\theta} \mid L) < R_{t_2}(\boldsymbol{\theta} \mid L)$; 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.

3. Objective Bayesian Procedures. As the authors (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 $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ 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*, $\{p(\boldsymbol{z} \mid \boldsymbol{\phi}), \pi(\boldsymbol{\phi} \mid \boldsymbol{\theta})\}$, where $\boldsymbol{\theta}$ would now be a hyperparameter vector, $p(\boldsymbol{z} \mid \boldsymbol{\theta}) = \int_\Phi p(\boldsymbol{z} \mid \boldsymbol{\phi}) \pi(\boldsymbol{\phi} \mid \boldsymbol{\theta})\, d\boldsymbol{\phi}$ would be the corresponding 'integrated' model, and a prior $\pi(\boldsymbol{\theta})$ would be required for the hyperparameter vector $\boldsymbol{\theta}$.

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 assumed model $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ 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 & 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).

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 ϕ, the consequences $L(M_i, \phi)$ 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,

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

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

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

closely related to cross-validation techniques; (for details, see Bernardo & 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 & 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 the authors 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 $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ 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 (1) 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 in 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 & Rueda (2002) and Bernardo & Ju árez (2003).

6. How possible and fruitful is it 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 & Savage, 1955; Bernardo & 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 $\overline{U}(a_1 \mid \boldsymbol{z})$ and $\overline{U}(a_2 \mid \boldsymbol{z})$, 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 $\gamma\,\overline{U}(a_1 \mid \boldsymbol{z}) + (1-\gamma)\overline{U}(a_2 \mid \boldsymbol{z})$, so that randomization for a single decision maker could apparently only be optimal if $\overline{U}(a_1 \mid \boldsymbol{z}) = \overline{U}(a_2 \mid \boldsymbol{z})$ 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 & Kadane (1997).

10. Is the formulation of personalistic probability by de Finetti and Savage the wrong way round? The authors (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. The authors 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 be relevant to *calibrate* the procedure (see the section on Foundations); 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 the authors 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, the authors have unfortunately chosen to ignore over 30 years of 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 & Smith (1994, Ch. 5); for critical overviews of the topic, see Kass & Wasserman (1996), Bernardo (1997), references therein and

ensuing discussion.

Reference analysis has been mentioned in the section on foundations 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 parameters 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 & Bernardo (1992); for a simple introduction see Bernardo & 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 & 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).

Acknowledgement

Research funded with grants GV01-7 of the Generalitat Valenciana, and BNF2001-2889 of the DGICYT, Madrid, Spain.

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. & Bernardo, J.M. (1992). On the development of reference priors. *Bayesian Statistics 4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, pp. 35–60 (with discussion). Oxford: University Press.
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*, Eds. N.G. Polson and G.C. Tiao, (1995) pp. 229–263. Brookfield, VT: Edward Elgar.
Bernardo, J.M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference*, **65**, 159–189 (with discussion).
Bernardo, J.M. & Juárez, M. (2003). Intrinsic Estimation. *Bayesian Statistics 7*, Eds. J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, pp. 465–475. Oxford: University Press.
Bernardo, J.M. & 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. & Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.*, **70**, 351–372.
Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: Wiley.
Berry, S.M. & 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*, Eds. H.E. Kyburg and H.E. Smokler, pp. 93–158. New York: Dover.
Jaynes, E.T. (1976). Confidence intervals vs. Bayesian intervals. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Eds. W.L. Harper and C.A. Hooker, **2** pp. 175–257 (with discussion). Dordrecht: Reidel.
Hewitt, E. & Savage, L.J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, **80**, 470–501.
Kass, R.E. & 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. & 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*, Eds. C.E. Shannon, and W. Weaver, 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.

Response by Groningen Statisticians

O.J.W.F. Kardaun, D. Salomé, W. Schaafsma, A.G.M. Steerneman and J.C. Willems

Groningen University, The Netherlands.
contact E-mail address: otto.kardaun@ipp.mpg.de

We enjoyed reading the comments and several of the references cited therein. Both discussants started with a general perspective and continued by giving their answers to some (Schweder) or all (Bernardo) issues. Expressed in a somewhat simplified fashion, Schweder basically agrees to the

views put forward by us, whereas Bernardo criticises our acceptance of (some) lack of probabilistic coherency. We first respond to Schweder and then to Bernardo in our reply below, which is the order in which the reactions were received.

At the beginning of his discussion, *Tore Schweder* clearly expressed the purpose of the paper. Many people may find it indeed 'stupid' to waste time on 'philosophical' questions that cannot be entirely resolved. However, more moderately expressed than by Dieudonné's flamboyant sentence in [*ii*] (indicating that open problems always much stimulated progress in algebraic geometry and number theory), as statisticians we find it both useful and attractive to discuss such critical issues in order to gain a better understanding of our work in daily practice.

We value Schweder's proposal in his answer to Q7 ($\equiv$ Question 7) to use the term *confidence distribution*, attributed to Efron [*iii*], for the (epistemic) fiducial probability distributions broached by Fisher, which is an improvement over the term inferential distribution we ventured to use in our answer to Q12. The reason for making this semantic distinction is that epistemic probability distributions cannot, in general, be described by sigma-additive measures. This has been illustrated in our answers to Q8, Q10, Q12 and Q14, and corresponds well with Schweder & Hjort's (2002) paper containing an application to whale fishery protocols. Whether we like it or not, as statisticians we are forced by reality to accept some form of lack of probabilistic coherency, since the axioms of probability theory are simply too restrictive if they are applied in general to epistemic probabilities.

It is apparent from Schweder's discussion of Q13 that conflicts appear if inferences have to be made about (non-linear) functions of the mean μ of a multivariate distribution, given the outcome x of $X \sim N_p(x, I_p)$ ($p \geq 2$). (In [80] something similar has been indicated even for the case $p = 1$.) Note that Schweder's discussion is related to Stein's phenomenon [135]. In our view it is more important to see this problem clearly than to give a gloss to it. The consequence is that probability (as a sigma-additive measure) 'continues to be the best scheme' in a certain number of situations only, which modifies Haavelmo's slightly more general plea described in Schweder's answer to Q3.

We basically agree with Schweder's answer to Q5, but we would like to add the comment that 'not making a decision' can well be one of the categorised options in decision theory, which relaxes the concept of a forced decision.

With respect to Q11 we remark that in an area such as plasma physics an 'essential understanding of what is actually happening' is perhaps not more easily attained than empirical predictions based on several collaborative, experimental investigations. The precise meaning of the word 'understanding' is sometimes not better defined than that of probability in statistics, however, in both cases leaving room to the prudent man in the sense of Aristotle for shaping the context of the discussion in practical situations to some extent.

While 'sidestepping the issue' in Q7 has led to a most useful suggestion by Schweder (in the vein of the Significa Movement), in Q13 the immediate sidestep may have prevented a fuller intrinsic reaction on the issue at stake. Therefore, we bring into recollection the interdependence between our answers, in particular between Q10, Q13 and Q14, and their more general predecessors Q6 and Q7. We appreciate Schweder's cautionary comment in the discussion of Q13 on the limitation of the objective Bayesian approach with respect to distributional inference in multidimensional situations.

José Bernardo gave a clear exposition of the approach in objective Bayesian decision theory which he qualifies, with a dash of justification, as being 'very different' from that of us.

We have three remarks to his introduction. First, we thought it should be obvious that the complete class theorem [*vii*], which has found its way in most standard texts on mathematical statistics [45] and which was mentioned in our answer to Q12, only provides an (*almost*) *necessary* but by no means a *sufficient* requirement for a procedure to be recommendable. The class of admissible procedures usually being very large, the real problem consists in selecting among them 'good' or even 'optimal' procedures in certain (well-described) frames of discussion, a topic analysed in [*v*, *vi*]. In this context, we have stated in Q12 'the choice of a prior distribution is awkward', which of course implicitly

admits the possibility of the existence of a prior distribution (at least so 'in the mind of the beholder'), in contrast to the more extreme frequentist viewpoint 'there is no prior distribution' expressed by Bernardo at the end of his answer to Q7.

Secondly, we admit that we did not stress in our answers the difference between the formal ('rationalist') and subjective ('personalist') Bayesian schools. This omission has occurred primarily since the (rather marked) difference was assumed to be a distinction the reader is familiar with, e.g. from [36, 37, 38]. However, we realize that thereby we may have fallen into the trap noted by O'Hagan in the discussion of Walley's paper which defies the Bayesian paradigm using a bag of marbles, see [145, p. 35]. We apologise for the possible ambiguity and use the opportunity to recover by mentioning that the answers to Q10, Q13, and Q14 pertain to the subjective approach (except in Q13 for the invariant prior $\pi(\theta) = c/\theta$ and the multidimensional prior based on Jeffreys's principle), whereas the answers to Q2 and Q12 pertain to both the subjective and the objective Bayesian approach.

Thirdly, it is also true that the real differences must not be exaggerated needlessly. A compromise approach between the various schools has been described in [80]. A remaining point of difference with the view described by Bernardo is that we do not restrict ourselves to a formal Bayesian school because the reference priors suggested in this approach are not regarded as compelling. Alternative principles such as 'strong unbiasedness' and 'strong similarity' have been proposed instead in [80], which lead to some revival, in a restricted context, of Fisher's fiducial approach.

For the sake of conciseness, we will not discuss all points in Bernardo's reply on the cryptic issues, but concentrate on a few important items only.

With respect to Bernardo's reply to Q7: Certainly in the context of public discussion we acclaim Bernardo's endorsement of the idea to use another word than probability, for instance 'credence' for a (rationalist or subjectivist) degree of belief. We do not follow him in his subsequent argumentation for retaining the same word 'probability' in all cases, and illustrate this as follows. In mathematical-logical discourse, 'sets' and 'propositions' both satisfy the axioms of a Boolean algebra, under the usual, obvious identification of operations, such as intersection with 'and', union with 'or', etc. Nevertheless, it makes sense not to confound the words 'set' and 'proposition', because their *intrinsic meaning* is essentially different, despite their identical axiomatic algebraic properties. The same state of affairs applies to frequency-based ('physical') and not necessarily frequency-based ('epistemic') probabilities, even if one holds the viewpoint that the latter 'should' satisfy the axiomatic properties of a finite, sigma-additive measure as formulated by Kolmogorov for physical probabilities, a viewpoint which we do not share, since we do not consider it to be a realistic model in a number of circumstances, as should be transparent from most of our answers, in particular those to Q8, Q10, Q13, and Q14.

In the second part of his reply, Bernardo's answer to Q7 refers to De Finetti's penetrating Representation Theorem, which utilises the notion of exchangeability, a generalisation of the concept of independent and identically distributed random variables, see [*iv*, *i*]. This generalisation, which amounts to permutational invariance of the multivariate distribution of a random sequence (say of Bernoulli experiments $X_1, X_2, \ldots$), leads to

$$P(S = s) = \int \binom{n}{s} \theta^s (1-\theta)^{n-s} d\nu(\theta)$$

for some measure ν rather than to

$$P(S = s) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

for a fixed θ, where the measure ν is interpreted as an epistemic, prior probability for P. However, the derivation of the representation theorem requires either finite or sigma-finite additivity of the multivariate distribution of $X_1, X_2, \ldots$, see e.g. [*iv*]. The assumption of sigma-additivity, 'derived' from a coherency principle and postulated to be necessarily adhered to by every 'rational' man, is

equivalent to the Kolmogorov axioms. The 'Munchhausianism' involved in this rationalisation is that one starts, a priori, from a principle equivalent to a (sigma-)additive measure satisfying Kolmogorov's axioms on a multidimensional outcome space, and then derives, under the model of exchangeability, the 'existence' of a prior (sigma-)additive measure ν in one dimension. The fallacy of 'misplaced prolific deductivism' in this case is that, subsequently, the one-dimensional prior measure ('since it has been derived') is considered to be 'generally applicable' to express uncertainties, degrees of belief ('credences') etc.

In our view it is good statistical practice not to use the model of assigning numerical, sigma-additive probabilities to express 'prior (lack of) information', unless this is done in a context where this is sufficiently warranted. Evidently, this often leaves room for statistical judgement. In that sense the statistical profession bears some resemblance with an art. Non-controversial examples are provided by empirical Bayes situations, and a plausible situation was also considered in Bayes's original physical example of repeatedly rolling a ball [16].

With regard to Q8 on axiomatic formulations in statistics, the viewpoint in our answer is less outspoken than Bernardo's, since we have indicated the need for the deductive framework to describe indeed adequately the class of physical phenomena it is intended for. To further clarify Fisher's viewpoint on this issue, which happens to be not too remote from our own, we present a quotation of a sentence from his paper on the logic of inductive inference [47]:

> "This, of course, is no reason against the use of absolutely precise forms of statement when these are available. It is only a warning to those who may be tempted to think that the particular code of mathematical statements in which they have been drilled at College is a substitute for the use of reasoning powers, which mankind has probably possessed since prehistoric times, and in which, as the history of probability shows, the process of codification is still incomplete."

Some form of Munchhausianism is (at times) lurking behind invariance considerations. This has been illustrated by the example of the invariant prior $\pi(\theta) = c/\theta$ in our answer to Q13. A somewhat more general type of such consideration is: 'If a decision problem is invariant under a group which acts transitively on the parameter space, then a best invariant procedure will exist.' Both Bayesian and non-Bayesian statisticians cherish such situations. The fallacy involved, however, is that one starts from the very existence of a 'unique', most appropriate rule. Next, one argues that a transformed situation is isomorphic to the original one and that, hence, the same procedure 'should be' implemented for the transformed problem. This procedure is consequently invariant under the group of transformations. The 'conclusion' is then that, obviously, a unique invariant, most appropriate rule exists. In this respect, we must concede that the defects arising from deductive reasoning by (over-) rationalisation are somewhat more deeply hidden than the more obvious ones which were expressed by Bernardo's car-salesman in Q10.

With respect to the last paragraph in Bernardo's answer to Q11, we notice that the fact that formal Bayesian methods tend to fail producing adequate results in multidimensional situations (Stein's example, which was also considered in Schweder's answer to Q13), demurs the (operational) validity of any derivation of these methods which is independent of the dimensionality of the situation. We agree with Bernardo that the responsibility for inferences that are statistically sound cannot be relegated to computations solely, independent of whether they are of a frequentist or of some other nature.

We admit that, because of the location of a singularity for $x + \alpha = 0.5$, the practical numerical approximation to the posterior median of the $Be(\alpha, \beta)$ prior distribution in Q13 is quite cumbersome to derive without appropriate basic computer software. Nevertheless, we go even further than Bernardo in his remark by stating that the combination of arguments in our answer to Q13 could (and probably should) have been formulated already between 1935 and 1939, i.e. after Fisher's paper

on inductive inference [47], and before the outbreak of the war, which unfortunately led Fisher and Neyman to embark on more immediate tasks [118].

Finally, we appreciate in his answer to Q14 Bernardo's concession about upper and lower probabilities being an interesting alternative ('important player') to Bayesian probability theory in a number of cases.

We much appreciate the reaction of both discussants and hope that the discussion may have contributed to a further clarification of at least some of the issues.

References

[*i*] Cifarelli, D.M. & Regazzini, E. (1996). De Finetti's contribution to probability and statistics. *Statistical Science*, **11**, 253–282.

[*ii*] Dieudonné, J. (1972). The historical development of algebraic geometry. *The American Mathematical Monthly*, **79**, 827–866.

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

[*iv*] Ressel, P. (1985). De Finetti-type theorems: an analytic approach. *Annals of Probability*, **13**, 898–922.

[*v*] Schaafsma, W. (1971). Testing statistical hypotheses concerning the expectation of two independent normals, both with variance one. *Indagationes Mathematicae*, **33**, 86–105.

[*vi*] Snijders, T.A.B. (1979). *Asymptotic Optimality Theory for Testing Problems with Restricted Alternatives.* Ph.D. Thesis, Groningen University. Also: MC Tract No. 113. Amsterdam: Centrum voor Wiskunde en Informatica.

[*vii*] Wald, A. (1947). An essentially complete class of admissible decision functions. *Annals of Mathematical Statistics*, **18**, 549–555.

Rejoinder by D.R. Cox

D.R. Cox

Nuffield College, Oxford, UK

At the risk of slight repetition I must explain why I had what may seem like the impertinence of posing these questions so concisely and without giving my own answers. It had been stressed to me that a lecture I was invited to give in Groningen must be for nonspecialists. At the same time it seemed likely that there would be quite a few statisticians in the audience and it was only fair to offer them something: a small piece of paper with a few questions appeared like a good idea. I did not expect a response but was very happy to get almost immediately a thoughtful letter from two students. In a sense out of this has grown the material printed above, including the valuable comments by Professors Schweder and Bernardo.

My own view is eclectic, of wishing to take what is valuable from different viewpoints. Of course this can degenerate into a vaguely benevolent intellectual cowardice but the defence is to regard the ultimate test to be that of relevance and fruitfulness in applications. The extraordinarily rich variety of applications of statistics then points against any single approach.

While my views are in most respects rather close to what may be called the Groningen school of statistics, I would have expressed them somewhat differently, on some occasions no doubt a bit more superficially.

[*Received October 2001, accepted November 2002*]