The Statistician 32 (1983)

Theory and Practice of Bayesian Statistics

DENNIS V. LINDLEY

2 Periton Lane, Minehead TA24 8AQ

1 Introduction

Good theory and good practice go hand in hand. Theory that does not have practical application is under-developed. Practice that is not supported by a sound theory is often subjective, misleading and inconsistent. My task here is to discuss the theory that complements the topic of this conference, with particular emphasis on the implications of the theory in practical situations.

2 Uncertainty and probability

Bayesian statistics is based on one, simple idea: *the only satisfactory description of uncertainty is by means of probability.* We are, all of us, surrounded by uncertainty: it plays a dominant role in all our lives. The Bayesian paradigm provides, in probability, a powerful tool for understanding, manipulating and controlling this pervasive, and often unpleasant, feature of our appreciation of our environment. The practical import is immediate: any unknown quantity should be described probabilistically. A more formal description of the paradigm follows.

You have a quantity, or a set of quantities, θ which is of interest to you but whose value is unknown to you. There are available some data D bearing on the uncertain value. (Notice that D, unlike θ, is known.) In addition you possess background knowledge pertaining to the situation under study. Denote this by H (for history?). The Bayesian view says that the appropriate description of your knowledge of θ in the presence of D and H is by the probability of θ, given D and H; and written $p(\theta|D, H)$.

3 Coherence

At this conference it is not necessary to spend much time on theoretical issues but at least we ought to recognize the reason for the description of uncertainty through probability, and in no other way. The reason is that the only way different judgements of uncertainty can fit together satisfactorily is to do so in exactly the same way as probabilities. If we progress beyond *single* statements of uncertainty and consider *sets* of statements, and if these are not to contradict each other, probability results. This basic idea is called *coherence*. There are two approaches: one based on decision-making, due to Ramsey (1926) and Savage (1954), and admirably discussed by DeGroot (1970); and another based on scoring rules due to de Finetti (1974a).

In passing, notice that coherence is not discussed in non-Bayesian statistics: a significance test at one sample size does not cohere with the same test at a different size. We need a name for non-Bayesian statistics. I propose calling them *Berkeley statistics* for two reasons: the campus of the University of California at Berkeley has one of the best departments of the non-Bayesian type in the world; and Bishop Berkeley (after whom the campus is named) was much criticized by the Rev. Bayes (for his views on Newton, see Holland (1962)).

It is not necessary to rehearse before this audience the rules by which probabilities cohere. The convexity (lying in the unit interval with zero for impossibility) addition and multiplication laws are the basic ones from which all others derive. According to the theory, all uncertainty judgements cohere this way. Furthermore, these are the *only* rules: any others can be derived from these three. Probability is the *only* tool for the study of uncertainty. (Some dissent from this view is mentioned in §18.)

4 Recipe

It is vitally important when considering practical applications to recognize that the Bayesian paradigm provides rules of procedure to be followed. I like to think of it as providing a *recipe*: a set of rules for attaining the final product. The recipe goes like this. What is uncertain and of interest to you? Call it θ. What do you know? Call it D, specific† to the problem, and H, general. Then calculate $p(\theta|D, H)$. How? Using the rules of probability, nothing more, nothing less. Examples are given in §§11, 19.

5 Decision-making

We have said probability is the only tool: this is correct whilst we take a *passive* view of the world in which we are content to describe our uncertainties. This procedure is called *inference*. To a Bayesian, inference is the numerical expression of uncertainties. Extensions of probability ideas are needed when we pass beyond this passive role and consider *action* in an uncertain world. Action will lead to consequences and the worth of a consequence is described by its *utility*. However, utility is probability-based, for if a sure consequence is replaced by an equivalent gamble on two reference consequences, one good and one bad, the utility of the sure consequence is the probability of the good consequence in the gamble. This immediately leads, by the extension rule (§8), to choosing that action which *maximizes expected utility* (MEU), the expectation being with respect to the probabilities already assessed in the inference. Thus, following Ramsey (1926), inference is that procedure which is needed for *any* decision problem (concerning θ) and can therefore be performed without a specific decision problem in mind.

We therefore see that the Bayesian paradigm is not merely a way of understanding the world, it also provides a way of controlling it. It is not just science but also technology.

6 Bayes rule

The basic rules of probability (§3) lead to other rules, of which the most famous is Bayes rule that gives its name to the subject. In the notation of §2, it says

$$p(\theta|D, H) \propto p(D|\theta, H)\, p(\theta|H) \tag{6.1}$$

† In some situations D may be vacuous; there are no data (see §15). The recipe still applies. H is never vacuous.

where all probabilities are functions of θ for fixed D and H. An essential feature of scientific method is the collection of data D, preferably by controlled experimentation, or alternatively by observation. Bayes rule is the essential accompaniment to this scientific activity, telling you, the scientist, how to revise your opinion of θ on observing D.

7 Likelihood principle

The only contribution the data makes in Bayes rule (6.1) is through the probability $p(D|\theta, H)$ considered as a function of θ called the *likelihood* (function). We have the important *likelihood principle* that the totality of information about θ provided by D is given by the likelihood of θ for the observed D. Here is an example of the principle.

Let θ take two values, G and $\bar{G}$, corresponding to guilt and innocence respectively of a defendant in a court case. Let D be new evidence being presented and H previous evidence. Then, by the principle, all the court requires from the new evidence is the likelihood of guilt, given all the evidence: in symbols $p(D|G, H)$ and $p(D|\bar{G}, H)$. Often D and H are independent given G, and given $\bar{G}$, so that we need only consider the probability of the new evidence first on the supposition of guilt, second on that of innocence. A forensic scientist (or any expert witness) is often puzzled by what he should say in court. With the likelihood principle the answer is clear: state what the evidence is and what its probabilities are under the two possibilities. In particular, he should *not* make probability statements about G, that is the court's prerogative. The point has been discussed by Evett (1982).

Generally, the principle requires consideration of a unique D, that observed, but all possible values of θ. It is in violating the principle that Berkeley methods typically come to grief. In considering data values that might have occurred but did not, as with a tail-area, significance test, they become incoherent.

The principle, whilst extremely important, is not pervasive: it is confined to the appreciation of data. For example, the whole range of possible values of D may be needed in order to write down the likelihood function. A Poisson distribution with zero-class missing has a different likelihood from the complete Poisson. Also in the design of experiments all possible data values need to be contemplated. It is only *after* the data are to hand that the principle applies.

8 Extension rule

A second, useful rule of probability that easily follows from the basic rules has no generally accepted name; I propose to follow Tribus (1969) and call it the *extension* rule. It says†

$$p(D|\theta, H)=\sum_{\phi} p(D|\theta, \phi, H)\, p(\phi|\theta, H)$$

and extends "the conversation" from θ to include ϕ. Its usefulness lies in the fact that the judgements about D often involve not merely the quantity of interest θ but also *nuisance* quantities ϕ. The rule allows these to be eliminated by summation (or integration). It is one of the more difficult problems of the Berkeley school to eliminate nuisance parameters: the Bayesian view has the single, universal method of integration. Alternative methods may lead to incoherence.

Bayesian methods include both Bayes rule (and hence the likelihood principle) and the extension rule. Methods based solely on likelihood (Edwards, 1972), are defective in that they have no general way of eliminating nuisance parameters.

† For simplicity of exposition, quantities are supposed to assume only a *finite* number of values.

9 Subjective probability

So far the discussion has been confined to the *calculus* of probability: in applications it is also necessary to consider its *interpretation*. There need only be one: a person's judgement about a quantity that is unknown to him. This is the *subjective* (or personalistic) view of probability. In this view probability does not exist outside the subject: there is no true probability but rather an expression of a relationship between you and the world. The word "subjective" is unfortunate because the Bayesian view is not subjective in the sense that data analysis is, where all manipulations are open to consideration and judgements are much at the whim of the analyst.† The Bayesian subject is severely constrained by coherence and by the inexorable role of data. As de Finetti (1974a, §1.4.1) has said, the only objective view of probability is the subjective one, because it can be tested by the rules it must obey. In practice, two scientists may disagree, but data and coherence will bring them together.

10 Exchangeability

Subjective probability, in its interpretation, contains no element of repetition: it has no frequency basis. Some statisticians confine themselves to situations admitting a frequency element. In Bayesian language they are restricting themselves to exchangeable (or partially exchangeable) cases, where the notion of *chance* arises through the operation of de Finetti's representation theorem (1975, §11.4.2). This is a severe and unnecessary restriction, for the paradigm equally applies to the unique occasion. The court case (§7) gives an example. Indeed, almost all situations ultimately call for a judgement about a *unique* occasion and it is a great strength of the Bayesian view that it can handle them. (And a weakness of the frequentist view that it cannot.) The point is of such practical importance that an example seems advisable.

11 Prediction

Suppose the data consist, in frequentist language, of a random sample from a population. In Bayesian terminology $D=x^{(n)}=(x_1, x_2, \ldots x_n)$ where, given the parameter θ, the x's are i.i.d. Consider another member of the population whose value x_{n+1} is unknown. Our uncertainty is, by the recipe, given‡ by $p(x_{n+1}|x^{(n)})$. By the rules of probability this is $p(x^{(n+1)})/p(x^{(n)})$, and the denominator is, on extending the conversation to include θ,

$$p(x^{(n)})=\sum_\theta p(x^{(n)}|\theta)\,p(\theta)=\sum_\theta \prod_i p(x_i|\theta)p(\theta)$$

by the assumed independence. The numerator has n increased by one. An alternative calculation gives

$$p(x_{n+1}|x^{(n)})=\sum_\theta p(x_{n+1}|\theta)\,p(\theta|x^{(n)}) \tag{11.1}$$

using independence again, and Bayes rule enables $p(\theta|x^{(n)})$ to be calculated.

The practical importance of these calculations can be seen by thinking of yourself as $(n+1)$ uncertain about your recovery (x_{n+1}) from a medical condition, when you have data on n patients in a medical trial. Provided you judge yourself to be exchangeable with these patients, the calculations apply. Notice the difference of roles played by $p(\theta|x^{(n)})$ and

† "the estimate would always depend ... on the *idiosyncracies* of the statistician" (Huber (1982), my italics).

‡ For simplicity, H is omitted from the notation.

$p(x_{n+1}|x^{(n)})$. The former provides a summary of the trial, the latter is specific to you. In a sense, the first is a scientific answer, the second a technological one. We return to this again in discussing coherence (§12). The emphasis on prediction of x_{n+1} instead of estimation of a parameter has been discussed by Geisser (1980) and elsewhere.

The Berkeley view does not easily accommodate opinions about *your* recovery. Tolerance intervals (an extension of confidence intervals) are available but are cumbersome. Notice that a confidence statement, although probabilistic, is not about θ (or x_{n+1}) but about $D=x^{(n)}$.

There is no space to discuss the topic here but notice how the above discussion of $x^{(n)}$ extends from the exchangeable case to where it constitutes a time series.

12 Coherence in practice

Coherence, underlying the Bayesian paradigm, may appear a rather theoretical concept. In fact, it is also extremely practical. Knowledge in science, comes from several sources: a worker in one country provides part of the answer, his overseas colleague another. Advances often occur through linkages between pieces of knowledge previously unconnected. Lawyers use precedence (coherence with earlier cases). These are all situations where coherence may be vital because it is *the* linking mechanism.

A simple example is provided by two experiments to measure a quantity θ: one gives Poisson counts over a time T with mean $T\theta$; the other observes times to failure, exponential with mean θ^{-1}. Two applications of Bayes theorem easily give an answer. The Berkeley school, with its two sample spaces, one discrete, the other continuous, finds the analysis cumbersome (Cox, 1980, p. 285).

Prediction analysis (§11) provides another example. There you $(n+1)$ were supposed exchangeable with the n patients. But the connection, or coherence, could have been expressed in other ways. If the trial had been conducted in the United States but you were in Britain; whilst you might consider the results relevant, you may feel θ is not the same on both sides of the Atlantic. You may express this as a relation between $\theta=\theta_1$ for the trial and θ_2 in Britain, plus one between you and θ_2. Then, with independence assumptions,

$$p(x_{n+1}|x^{(n)})=\int\int p(x_{n+1}|\theta_2)\,p(\theta_2|\theta_1)\,p(\theta_1|x^{(n)})\,d\theta_1\,d\theta_2$$

There are other possibilities.

Coherence has a broader sweep. In sampling inspection the methods usually employ a standard likelihood; for example, normal or binomial. Adequate data storage of the many examples of inspection might enable us to see whether these likelihoods are appropriate, or whether others are more appropriate.

To widen the context even further, we now know that the exponential family does not behave in some situations in quite the way we might find reasonable. For example, it effects a compromise between prior and likelihood that often seems forced. Distributions with longer tails behave differently (O'Hagan, 1979). Coherence suggests using such distributions; a possibility which can be turned into a reality with the Bayesian paradigm that does not strongly depend on the concept of sufficiency. A one-way layout with a t-distribution is easily possible for a Bayesian. Our only straitjacket is the principle of coherence, not of technique (§14).

A more extreme case of coherence is provided by examples of general, scientific experience; as when it is observed that empirical relations between univariate quantities are often well described by polynomials of low degree. Does not this mean that even in a novel problem involving such a relationship we might have prior opinions that the higher-degree

coefficients are smaller than those of lower degree? I see coherence as one of the major tools of future scientific studies.

13 Completeness of a probability statement

In describing the theory it has been emphasized (§3) that *only* probability is required. The practical relevance of this remark is apparent when the Bayesian and Berkeley views are contrasted. The Bayesian only requires $p(\theta|D)$ – we had an example in §11 where even when passing from $D=x^{(n)}$ to x_{n+1} the data contributed only $p(\theta|x^{(n)})$ in (11.1). The probability distribution is a *complete* statement and point or interval estimates, or significance tests, are not necessary. "No estimation problem *per se* is acknowledged to exist" (de Finetti, 1961). Estimates may be calculated – for example, $E(\theta|D)$ is a possible, point estimate – but they are derived from the full specification and information is necessarily lost in using them. They may help in appreciating the distribution. And let us note in passing that one of the most challenging of the many, difficult, technical problems in Bayesian statistics is to find ways of appreciating distributions when θ has high dimensionality.

14 Standpoint and techniques

This reduced need for many of the tools of standard statistical practice is especially hard for someone who has been trained in that practice to understand. Indeed, one of the hardest things for a Bayesian to do is to cast off the frequentist shackles. This has been forcibly discussed by de Finetti (1974b) in his distinction between the *Bayesian standpoint* and *Bayesian techniques*. The former is what we have been discussing. The Berkeley school uses the latter because the Bayes solutions essentially provide all admissible techniques. We must not be entrapped by the techniques and mindlessly pursue a standard procedure. The Bayesian standpoint always requires thinking about the problem in its real terms; forgetting about the Greek letters and concentrating on reality. Here is a simple, possibly over-simple, example to illustrate the point.

In n Bernoulli trials with r successes, standard theory tells us that r/n is a good estimate of the chance θ of success in a single trial. Contrast a situation in which each trial is a single, human birth and success is a girl, with one in which each trial is an inspection of a leaf and success is freedom from viral infection. Whilst r/n might be a reasonable value in the latter case, something near 0·49 is more reasonable in the former, for modest values of n, simply because H tells us so much about human sex ratios. Frequentists have difficulty in distinguishing between sex and viruses.

15 Statements of variability

Another error of technique of a different type leads to an unsound statement of variability, or error. Scientists are often interested in determining the value of a constant – the acceleration due to gravity at a site, for example – and the standard, statistical technique is to consider the data to form a sample of size n from a normal distribution $N(\theta, \sigma^2)$. The estimate of error is $\sigma/\sqrt{n}$. This tacitly assumes that θ is the constant: but is this reasonably so? Surely not, for there is the possibility of bias in the measuring device. A Bayesian will need to consider this, possibly without the aid of data specific to the problem but using only scientific knowledge of this and similar measuring devices: H is strong, D is empty. If θ is $N(g, \tau^2)$ where g is the constant, the appropriate error is $(\tau^2+\sigma^2/n)^{1/2}$, which may be much larger than $\sigma/\sqrt{n}$. The scientific literature is full of examples of underestimation of error because only sampling error has been included (Stigler, 1977).

16 Flexibility of approach

The difference between standpoint and technique can be illustrated in another way. The Bayesian standpoint is rigid and well defined: it comprises only the rules of probability. All it demands is coherence but that it insists upon. By rigidly demanding so little it enables there to be tremendous flexibility in technique. (The Berkeley school is rigid in its technique.) This can be seen at various levels: the use of distributions outside the exponential family has already been mentioned (§12). At a basic level there is no requirement to do the probability calculations always in the same way. We usually think about $p(\theta|H)$ and $p(D|\theta, H)$, using Bayes theorem to find $p(\theta|D, H)$. But we could go directly to $p(\theta|D, H)$. Here we should have to check for coherence as D varies but no one probability has status different from that of another, so all approaches are permissible. A practical application, with agreed $p(D|\theta, H)$ might be to start with $p(\theta|D, H)$ and see what $p(\theta|H)$ could have generated it. Or again, why bring in an unobservable, θ, at all? It would be possible to argue entirely within $p(x_{n+1}|x^{(n)})$ (§11). Sturrock (1973) provides a novel approach when discussing the combination of theoretical and experimental knowledge.

17 Points of dispute – reference distribution

I have written as though the Bayesian paradigm was a unique, well-defined and agreed position. This is properly not so: there are important, undecided issues to consider. All agree on the probabilistic description, but not on the exact rules of probability, nor on their interpretation. Does the addition rule apply to an enumerable infinity of events, or just to a finite number (Hill, 1980)? Allied to this is the question of whether improper distributions – ones that are not finitely integrable – can be used. The paradoxes particularly associated with Stone (1976) illustrate the pitfalls. Good (1965) sees a need for many types of probability. These are not theoretical niceties that can be left to the mathematicians but lead to points of practical substance. This is seen in the extensive literature on "ignorance".

Attracted by the apparent objectivity of Berkeley statistics, many workers have sought for an objective "prior" $p(\theta|H)$, especially when H is almost vacuous so that there is little information about θ. The originator is Jeffreys (1961). The idea is that if we have a random sample $x^{(n)}$ from a distribution with density $p(x_i|\theta)$ then there is a natural probability for θ associated with it. An alternative view (Bernardo, 1979), is to regard this distribution for θ as a reference distribution from which the information in other distributions can be measured. Either way, we obtain a useful probabilistic description for θ which can be combined with the agreed likelihood and an "objective" analysis of the data is obtained. My personal view is opposed to this and I see the argument as an example of technique overriding standpoint (§14) in which the Greek letter takes precedence over its meaning. But the point is by no means settled and has some importance in the analysis of data.

18 Points of dispute – non-probabilistic techniques

Another point of discussion amongst Bayesians is whether the "only" aspect of the paradigm is correct: do we need techniques that go outside the strict, probabilistic argument used above? Dempster (1980), Good (1967) and, most recently and eloquently, Box (1980) have argued that tail-area, significance test of the Berkeley school have a role to play in inference (as explained above (§7) they fall outside the Bayesian method in their violation of the likelihood principle). The point arises because there are situations in which a "natural" hypothesis occurs, but alternatives are not easy to specify. (It is easy to say that two distributions

are the same; more difficult to say how they might differ.) Everyone agrees that it only makes sense to test a hypothesis against an alternative. A tail-area test involves alternatives in considering what criterion to use but only to a degree that is more practically realistic than the precise modelling required by the full Bayesian treatment. My difficulty with the approach is that *whatever* alternatives were to be considered, only the data actually observed – and not what might have occurred – are needed, so why use the might-have-beens at all? The difficulty is closely related to the need for models, more so in the Bayesian, than in Berkeley, techniques.

19 Models

The whole nature of a model is obscure. The Bayesian sees a model as a helpful way of specifying the probabilities that are essential to his method for studying uncertainty. But part of a statistician's job does not involve uncertainty. Statisticians spend some of their time studying data. The book by Anscombe (1981) provides many illustrations where the question of Bayes or Berkeley does not arise because appreciation of the data is all that is initially required. For a simpler example, a set of pairs (x, y) of values is available, which will be plotted in various ways with a view to understanding and simplifying the situation. At this point there is no role for the Bayesian argument, for the discussion concerns only the data, which is known. A simplification would consist in fitting a straight line to the data. This might appear to call for a Bayesian approach to the uncertain nature of the line but there is still no uncertainty present that refers to observables, only to parameters, and to pass from known values to unobservable constructs is a mighty leap. The problem becomes more realistic when uncertainty about observables arises: thus if we ask what would happen to y at another value of x. In an extension of the earlier notation of §11, we need $p(y_{n+1}|x^{(n+1)}, y^{(n)})$ and an inference is required.

The "trick" that a model provides is to simplify the complicated structure involving $2(n+1)$ quantities by introducing others, termed parameters. Paradoxically, like the extension rule, more makes for simpler. It is then that the straight line fit might be appropriate by modelling $y_i - \alpha - \beta x_i$ as exchangeable with parameter (α, β). We split the probability structure into data, given parameters, and parameters. This is a simple hierarchy that may be extended with parameters depending on hyperparameters and so on. We sometimes fit a line too mechanically, sacrificing standpoint for technique. Our task is always to describe the situation coherently (probabilistically).

One feature possessed by a model is best explained in the context of Bayes theorem (6.1). It is usually written

$$p(\theta|D, H) \propto p(D|\theta)\, p(\theta|H)$$

in which, given θ, D and H are assumed independent. In other words, the model separates D from H, with θ acting as the intermediary: or θ gives a Markov structure to the sequence (H, θ, D). With this admitted, Bayesians might agree on the model specification even if their histories differ.

20 Model probabilities

In the subjective view all probabilities are on the same footing (§16) so that that for a parameter differs in no essential way from that for an observable (although one is testable, the other not). Many people, perhaps because of their statistical upbringing, think of probabilities for exchangeable events as different from others: or confuse probabilities and

chances. The clearest illustration of the unity is provided by the fact that it is not always clear what goes into the likelihood (model) and what into the "prior". Thus, in Model I analysis of variance, the block constants, treated as parameters, have their probabilities described in the prior; whereas in Model II they are exchangeable and part of the likelihood. Yet the overall probability specifications may be the same for both models.

Another example is provided by calibration. In §19 it was suggested that we might judge $y_i - \alpha - \beta x_i$ to be exchangeable and this leads easily to an evaluation of $p(y_{n+1} | x^{(n+1)}, y^{(n)})$. In calibration, with the same exchangeability assumption for exact x and calibration measurement y, we require instead $p(x_{n+1} | x^{(n)}, y^{(n+1)})$. This inverse probability differs in no essential way from the direct one for y_{n+1}.

21 Models and theories

There is a distinction to be made between models and theories rather like that between tactics and strategy: one looks at the specific situation, the other has a more global view. The Bayesian can often see the difference as being between different levels of a hierarchy. To illustrate in the context of two quantities, y and x. A *theory* may describe a linear relationship between them with values (α, β) (§19). But when we come to a data set $(x^{(n)}, y^{(n)})$ it may be *modelled* with values (α_j, β_j) for the jth data set. Then there is a further need to model the dependence of the (α_j, β_j) on (α, β). The point has been well made by Ehrenberg (1968) though outside the Bayesian framework – and he speaks of laws rather than theories. He makes the important point that interest rests primarily on (α, β), the theory, rather than the almost incidental model values (α_j, β_j). The point is related to that made about biases in measuring a constant in a theory (§15).

22 Probability assessment

The most important practical problem in the implementation of the Bayesian paradigm is the determination of the numerical values for the probabilities: expressed loosely but emphatically, "where did you get that prior?" Berkeley statisticians say, correctly, that we are unable to do this and then, erroneously, infer the Bayesian method is useless.

Ramsey's discovery that the laws of probability govern uncertainty parallels Newton's discovery of the laws of motion; laws that required for their exploitation methods of measuring speeds, forces, etc. Ramsey's laws require us to measure probabilities. Physicists built apparatus to measure the quantities required for Newtonian mechanics, they did not sit back and say force cannot be measured, so Newtonian mechanics is useless. So ways have to be found to measure probabilities. We cannot expect the "prior" to come naturally any more than arithmetic does: we have to learn probability assessment. Even those who favour reference priors have to admit the problem because their methods fail without exchangeability.

Probability assessment is a topic to be studied jointly by statisticians and experimental psychologists. The latter have rather clearly demonstrated that people are not naturally Bayesian: Kahneman *et al.* (1982) provide an excellent account. Psychologists have carried out experiments to see how good people are at assessing probabilities. To my mind, interesting as their results are, the experiments suffer from the defect that the subjects have typically been untrained in the calculus of probability. It is as unreasonable to expect people to make a good job of probability assessment without training as it is to expect good arithmetical abilities without learning to multiply. (Incidentally, many who have been trained do not

fully appreciate the product rule because of the extensive use of independence concepts.) There are related problems in the determination of utilities.

23 Scoring rules

One way of training an informed person to make probability assessments is to use a scoring rule. The familiar quadratic rule gives a penalty score $(a-1)^2$ to a subject who assesses the probability of an event A as a when A is true, and score a^2 when A is false. The encouragement to minimize his total score has a salutory effect on his assessments: an over-confident person moves away from his 1's and 0's; an unsure one ceases to vacillate around $\frac{1}{2}$. Other rules are possible: for example, the logarithmic rule, $-\log a$ if A true and $-\log(1-a)$ if false. Is one rule better than another? Or is it wise to think in terms of probabilities? The rule with scores $e^{-(1/2)a}$ if A true and $e^{(1/2)a}$ if A false leads to a statement of log-odds. Perhaps, as in Bayes theorem, log-odds are easier to assess than probabilities: we do not know.

One danger that has to be considered is that of an *implicit* scoring rule, in contrast to an explicitly stated one. Here is an example. A subject is required to assess the median demand for a product during the next month. If the storage cost is low but the loss of a potential customer through unavailability of the product is high, he may tend to state a higher median than that which truly reflects his beliefs. The implicit score leads to a bias just as scores $2(a-1)^2$ if A true and a^2 if A false lead to an overstatement of probability. An explicit statement of the rule seems necessary if biases are to be avoided.

24 Coherent assessment

Scoring rules are often used in a situation in which the subject assesses the probabilities of several, unrelated events. This does not exploit coherence for he need only use the convexity rule. It seems essential for good probability assessment to use all three rules and the full force of coherence. A subject should be asked about related events. Thus in addition to event A, he might be asked to consider A, conditional on B and then on $\bar{B}$. If he is also asked about B, the extension rule (from A to B) provides a coherence check. The more related questions he is asked, the more opportunities there are for coherence. As has been argued elsewhere (Lindley *et al.*, 1979), the situation has analogies with surveying where the surveyor makes more measurements than are strictly necessary and uses the coherence of geometry, allied to least squares, to make good assessments. (The surveyor's measurements are mainly of angles, not lengths; which leads one to speculate again (§23) whether probabilities are the best expressions.)

We have previously emphasized (§16) that all probabilities are on a par and that consequently any may be revised if another implied one seems unsatisfactory. The longer-tailed distributions, like t, again (§12) provide an example. With $x \sim N(\theta, 1)$ as likelihood and $\theta \sim N(0, 1)$ as prior, we compromise on $\frac{1}{2}x$ as the posterior expectation of θ with variance $\frac{1}{2}$. The precision is spurious and absurd when x is 5. But a t-distribution will avoid the problem, giving a more sensible compromise and a larger variance. The normal distribution does not cohere with other views that we possess. Our subject is coherence: it must be the major tool in asesessment.

25 Conclusion

The Bayesian paradigm concerns uncertainty. Its only tool is coherence expressed through the three laws of probability. It applies to statistical, repetitive situations where a judgement

of exchangeability is possible. But it is also applied to unique situations. We are uncertain about the inflation rate next year, the world's oil reserves, or the possibility of nuclear accidents. All these can be handled by subjective probability. What marvellous practical possibilities this suggests.

References

Anscombe, F. J. (1981). *Computing in Statistical Science Through APL*. Springer-Verlag, New York.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B*, **41**, 113–28 (with discussion 128–47).

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, **143**, 383–404 (with discussion 404–30).

Cox, D. R. (1980). Local ancillarity. *Biometrika* **67**, 279–86.

De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, **1**, 199–210.

De Finetti, B. (1974a). *Theory of Probability*, Vol. I. Wiley, London (translated from 1970 Italian edition).

De Finetti, B. (1974b). Bayesianism: its unifying role for both the foundations and applications of statistics. *International Statistical Review* **42**, 117–30.

De Finetti, B. (1975). *Theory of Probability*, Vol. 2. Wiley, London (translated from 1970 Italian edition).

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Dempster, A. P. (1980). Bayesian inference in applied statistics. In *Bayesian Statistics* (ed. J. M. Bernardo *et al.*), pp. 255–79 (with discussion 279–91). University Press, Valencia.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press.

Ehrenberg, A. S. C. (1968). The elements of lawlike relationships. *Journal of the Royal Statistical Society A* **131**, 281–302 (with discussion 315–29).

Evett, I. W. (1982). "What is the probability that this blood came from that person?" A meaningful question? *Journal of the Forensic Science Society* (to appear).

Geisser, S. (1980). A predictivistic primer. In *Bayesian Analysis in Econometrics and Statistics* (ed. A. Zellner), pp. 363–81. North-Holland, Amsterdam.

Good, I. J. (1965). *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*. MIT Press, Harvard.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society B* **29**, 399–418 (with discussion 418–31).

Hill, B. M. (1980). On finite additivity, non-conglomerability and statistical paradoxes. In *Bayesian Statistics* (ed. J. M. Bernardo *et al.*), pp. 39–49 (with discussion 49–66). University Press, Valencia.

Holland, J. D. (1962). The reverend Thomas Bayes, F.R.S. (1702–61). *Journal of the Royal Statistical Society A* **125**, 451–61.

Huber, P. J. (1982). Current issues in robust statistics. In *Some Recent Advances in Statistics* (ed. J. T. de Olivera and B. Epstein), pp. 183–96. Academic Press, London.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edn. Clarendon Press, Oxford.

Kahneman, D., Slovic, P. and Tversky, A. (Eds). (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Lindley, D. V., Tversky, A. & Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society A* **142**, 146–62 (with discussion 162–80).

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society B* **41**, 358–67.

Ramsey, F. P. (1926). Truth and probability. Reprinted in *Studies in Subjective Probability* (1964). (eds H. E. Kyburg Jr and H. E. Smokler). Wiley, New York.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Stigler, S. M. (1977). Do robust estimators work with *real* data? *Annals of Statistics* **5**, 1055–78 (with discussion 1078–98).

Stone, M. (1976). Strong inconsistency from uniform priors. *Journal of the American Statistical Association* **71**, 114–16 (with comments 117–25).

Sturrock, P. A. (1973). Evaluation of astrophysical hypotheses. *Astrophysics Journal* **182**, 569–80.

Tribus, M. (1969). *Rational Descriptions, Decisions and Designs*. Pergamon, New York.