Statistical Science
1990, Vol. 5, No. 1, 44–89

The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics

Dennis V. Lindley

Abstract. The first five sections of the paper describe the Bayesian paradigm for statistics and its relationship with other attitudes towards inference. Section 1 outlines Wald's major contributions and explains how they omit the vital consideration of coherence. When this point is included the Bayesian view results, with the main difference that Waldean ideas require the concept of the sample space, whereas the Bayesian approach may dispense with it, using a probability distribution over parameter space instead. Section 2 relates statistical ideas to the problem of inference in science. Scientific inference is essentially the passage from observed, past data to unobserved, future data. The roles of models and theories in doing this are explored. The Bayesian view is that all this should be accomplished entirely within the calculus of probability and Section 3 justifies this choice by various axiom systems. The claim is made that this leads to a quite different paradigm from that of classical statistics and, in particular, problems in the latter paradigm cease to have importance within the other. Point estimation provides an illustration. Some counter-examples to the Bayesian view are discussed.

It is important that statistical conclusions should be usable in making decisions. Section 4 explains how the Bayesian view achieves this practicality by introducing utilities and the principle of maximizing expected utility. Practitioners are often unhappy with the ideas of basing inferences on one number, probability, or action on another, an expectation, so these points are considered and the methods justified. Section 5 discusses why the Bayesian viewpoint has not achieved the success that its logic suggests. Points discussed include the relationship between the inferences and the practical situation, for example with multiple comparisons; and the lack of the need to confine attention to normality or the exponential family. Its extensive use by nonstatisticians is documented.

The most important objection to the Bayesian view is that which rightly says that probabilities are hard to assess. Consequently Section 6 considers how this might be done and an attempt is made to appreciate how accurate formulae like the extension of the conversation, the product law and Bayes rule are in evaluating probabilities.

Key words and phrases: Bayesian statistics, sequential probability-ratio test, likelihood, likelihood principle, likelihood ratio, prior probability, posterior probability, sample space, loss function, utility, expectation, errors of the two kinds, risk function, ancillarity, nuisance parameters, coherence, inductive logic, inference, hypotheses, parameters, theories, models, large and small worlds, decision-making, scientific revolutions, paradigms, future observations, point estimation, counter-examples, consistent estimates, probability assessment, multiple comparisons, non-orthogonality, shrinkage estimates, generalized linear models, Kalman filter, credibility, clinical trials, least squares, extension of the conversation, Bayes rule, additive coherence, product law, log-odds.

Until his retirement in 1977, Dennis V. Lindley was Professor of Statistics and Head of the Department of Statistics and Computer Science at University College, London. His mailing address is 2 Periton Lane, Minehead, Somerset TA24 8AQ, England.

PREFACE

It is an honour to be asked and a pleasure to give these lectures to members of the Institute of Mathematical Statistics. The lectures are devoted to the foundations of inference and to showing how important the foundations are in determining statistical practice. Except in Section 6, little attention is paid to the necessary technical developments. This emphasis on foundations is partly due to my own interests and partly due to space limitations. It most certainly does not arise because of any lack of appreciation of the value of technicalities. Foundations, technicalities and practice are all important and our subject is dependent on all three.

The aim of these lectures is to explain Bayesian statistics as a new paradigm. In Section 1, Wald's work is contrasted with Bayesian, particular emphasis being placed on the role of the sample space. In Section 2, an attempt is made to describe the basic problem of statistics as part of scientific method, and in Section 3 to argue, and to explore the unappreciated consequence, that its solution must be Bayesian. Section 4 extends the discussion to decision-making, and Section 5 is a foundational glance at statistical practice. A final Section 6 contains an attempt to face up to the famous, legitimate objection to Bayesian statistics that the prior may be unknown. It is argued that the measurement of probabilities is a problem that has not been seriously considered, even by probabilists, and a beginning is made on its resolution.

There is no generally accepted name for non-Bayesian statistics. Since Bayes had little to contribute to Bayesian statistics, it is not inappropriate to refer to Berkeley statistics, since the two ecclesiastics disagreed during their lifetimes, and because the University of California campus named after the latter has perhaps the best department broadly holding to that view. To vary the style, the terms coherent and sampling-theory (due to Box and Tiao, 1973) will be used.

1. COHERENCE

1.1 Bayesian Ideas in Wald's Work

Wald was responsible for two major results, in addition to many other researches that we would all be proud to be able to call our own. The first was the invention of the sequential, probability-ratio test (SPRT) (Wald, 1947) and its optimality property (Wald and Wolfowitz, 1948). The second was the proof that the only admissible solutions to a decision problem are essentially Bayes solutions and that they form a minimally complete class (Wald, 1950). Both these results are Bayesian. The second obviously so. The first is because if l_i is the likelihood of the data under H_i ($i = 0, 1$) and π_i the prior probability of H_i, the SPRT, which consists in sampling as long as $A < l_0/l_1 < B$, where A and B are constants, and stopping as soon as either inequality is breached, may be rewritten in terms of posterior odds as $A' < \pi_0 l_0/\pi_1 l_1 < B'$; which is equivalent to stopping as soon as the posterior probability of H_0 is sufficiently large or small; in Bayesian terms, as soon as Your belief in one of the two hypotheses is sufficiently strong. An essential feature of the SPRT is that the critical values of the likelihood ratio (or of the odds ratio) do not depend on the number of the observations but are genuinely constants.

It is remarkable that Wald's work led, especially in the United States and most particularly by members of this Institute, to the development of a school of statistics that is resolutely anti-Bayesian and eschews the use of probability distributions over parameter space except as a tool to produce admissible decision procedures whose non-Bayesian properties are then studied.

1.2 The Incompleteness of Wald's Results

The two results just cited are incomplete (in a nontechnical sense) in that they only refer to the optimality of a *class* of procedures (SPRT, or Bayes) and do not, on their own, produce a *single* method that can be preferred to others in the class. Wald, and others, have written on the minimax procedure to select a unique optimum, but this has not found general favour because of the unsatisfactory results it gives even in simple situations like the estimation of a binomial parameter (Wald, 1950, page 142). A deeper objection to minimax will be given in Section 1.5. The most popular method of obtaining a unique procedure, well illustrated in the case of the SPRT, is to use other sample-space properties. Wald (1947) showed that, to a good approximation, the choice of $A = \alpha/(1 - \beta)$ and $B = (1 - \alpha)/\beta$ will lead to probabilities α and β of false acceptances of the hypotheses. Consequently, fixing α and β will determine A and B and hence the test. In more general situations the power curve can be similarly used. In other cases, restrictions on the class of procedures, for example to unbiased estimates, will often lead to a unique optimum.

(There is another aspect of Wald's work where the incompleteness remains, even today. It is all based on a loss function and the minimization of the expected loss over sample space: the risk function. Yet no satisfactory explanation appears to be available of what a loss function is or why only its expectation should be relevant. We return to this point in Sections 4.1, 4.2.)

1.3 The Use of Sample-Space Criteria

It is in this use of sample-space criteria to select a unique decision procedure out of a class that the

Waldean and the Bayesian part company. This is most easily seen in the case of two simple hypotheses, H_0 and H_1, either in the fixed-sample-size case or the SPRT. The former will use α and β, the errors of the two kinds, which involve integrations over sample space. The latter will use π_0, the prior probability of H_0 ($\pi_1 = 1 - \pi_0$), the cost of sampling and other losses.

With a good deal of generality, the situation can be described as follows. Let X be the sample space of points x and Θ the parameter space of points θ. These are connected by $p(x|\theta)$, the probability density of X for a given θ (with respect to some measure). Let D be the decision space of points d, and $L(d, \theta)$ the loss in selecting d when θ obtains. Waldean concepts use a decision function $\delta(x)$ from X to D, that prescribes what decisions to take when x is observed, and base selection of it on the risk function

$$(1) \qquad R(\delta, \theta) = \int_X L(\delta(x), \theta)p(x \mid \theta)\, dx.$$

The Bayesian approach uses a probability density $\pi(\theta)$ over parameter space and chooses as the optimum decision that d which minimizes the expected loss

$$(2) \qquad L^*(d, x) = \int_\Theta L(d, \theta)p(\theta \mid x)\, d\theta,$$

where $p(\theta \mid x)$ is the density of θ, given x, obtained by Bayes theorem. This analysis does not use the sample space except insofar as X may affect the likelihood $p(x \mid \theta)$ regarded as a function of θ for fixed x. Once x is observed and the likelihood available, X is irrelevant.

(A Bayesian can use the decision function $\delta(x)$ and minimize the average risk

$$(3) \qquad \int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta,$$

but a simple reversal of the orders of integration of θ and x easily reduces this to the simpler minimization of (2). Raiffa and Schlaifer (1961) refer to this as the normal (in the sense of "usual") method, and to (2) as the extensive method.)

It is in the contrasting use of two entirely different expectations, (1) and (2), that the disagreements between the two schools are most clearly displayed. The sampling-theorist objects to the Bayesian's use of an arbitrary $\pi(\theta)$. The latter objects to the former's use of an arbitrary sample space. When Berkeley says to Bayes, "where did you get that prior?", Bayes can respond with, "where did you get that sample space?"

Since the arbitrariness of the sample space is not often appreciated, it might be worth discussing it. The practical reality is the data x (not X), the parameter space Θ and the likelihood function $p(x \mid \cdot)$ for fixed x and variable θ. The sample space X is, to use Jeffreys' (1939) vivid description, the class of observations that might have been obtained but weren't. Both in practice and in theory, this class can be hard to specify. (The problem of experimental design is not being discussed.)

Let me digress to answer a point raised by two referees, and others privately, to the effect that the sample space X and its associated densities are the primary entities from which the likelihood is derived. This need not be so. Although it is customary for any paper in probability to begin with the triplet (X, A, p), the space, the sigma-algebra and the probability measure (or density) and, in statistics, to extend to a set of probabilities indexed by a parameter, this complete specification is not necessary and often extends beyond the bounds of the reality. Why, when discussing probabilities, is it necessary to have them defined for more sets than those of interest? Why insist on *all* members of the field? Why, when a thumb-tack has been tossed 6 times, with 4 of these resulting in the point facing upwards in an observed order, do we have to think of other possibilities? The likelihood is $\theta^4(1 - \theta)^2$ without the need for the extra considerations. If observations have been made of n normal quantities, the log-likelihood is $-n \log \sigma - \frac{1}{2} \sum (x_i - \theta)^2/\sigma^2$ irrespective of the fact that, had the observations been other than they were, sampling would have stopped before n. The (X, A, p)-introduction is a useful starting position for many problems but not when the data are to hand. Then the likelihood function is primary and, as we shall try to show, the sample space is an arbitrary addition imposed on it.

1.4 Difficulties in the Definition of the Sample Space

A common case is the observation of a random sample of size n and for the statistician to take the sample space to be all samples of that size. But it often happens that the scientist had arrived at n by a random procedure: some of the plants may have died; time or money may have run out. Were the experiment to be repeated—a concept uppermost in the frequentist's mind—a different n might result. By what reasoning can the statistician justify fixing n to provide the sample space? That fixing n can affect the ultimate choice of decision function is seen by contrasting the effects of positive or negative binomial sampling on the unbiased estimate of the chance parameter. On the rare occasions when some justification is attempted, an appeal is often made to ancillarity, but Basu (1964) has shown that this is unsatisfactory because there are ancillaries that are indefensible.

As an example, let x be uniformly distributed in $[\theta, 1 + \theta)$. Then the fractional part of x is ancillary but x, conditional on this ancillary, has a one-point distribution providing a sample space that is too restricted. The reduction of the sample space must rest on more than ancillarity—or be an incorrect procedure.

There are many examples of constructed sample spaces that differ from the reality of the experiment. With a contingency table, the margins are often supposed fixed. In studying the bivariate regression of y on x, it is usual to hold the x's fixed at their observed values. If the parameterization is in terms of (θ, λ), where θ is of interest and λ is a nuisance parameter, we can write, for a statistic $t(x)$,

$$(4) \quad p(x \mid \theta, \lambda) = p(t(x) \mid \theta, \lambda)p(x \mid t(x), \theta, \lambda).$$

If λ is absent from either of the two factors on the right-hand side, that factor alone may be used for inference about θ and the other discarded. This has led to many varieties of likelihood and, by implication, unusual sample spaces. There is no doubt that many of these procedures are useful. My point is that their theoretical underpinning is weak.

So Berkeley uses a sample space; Bayes employs a prior. Both have elements of arbitrariness. It is possible to avoid both by using an approach based purely on the likelihood function, but this has difficulties in handling nuisance parameters and appears to be ill-adapted to decision-making. So the contrast remains. I now argue that Wald's approach had an essential ingredient missing. Once that is inserted, the Bayesian viewpoint is seen to be preferable.

1.5 Coherence

The missing ingredient is most easily appreciated in the work of Neyman and Pearson that Wald was later to generalize. Still considering the case of two, simple hypotheses, they advocated fixing α, the error of the first kind, and minimizing β, that of the second. But they never considered whether doing this, both for a sample of size n_1 and for a sample of a different size n_2, made sense. In general, consider two problems sharing the same parameter and decision spaces but differing in their sample spaces. A practical example is that of two scientists performing different experiments to investigate the same phenomenon. The key question is: do the statistical analyses of the two problems fit together, or cohere? In particular, how would the analyses compare with the single analysis of the two experiments when combined? It is the concept of coherence that is absent from sampling-theory statistics.

To investigate this further, let the parameter and decision spaces both have two elements, (θ_0, θ_1) and (d_0, d_1) respectively. Suppose that, in some sense, d_i is correct if θ_i obtains $(i = 0, 1)$, and let this be reflected in a loss function with $L(d_i, \theta_i) = 0$ and $L(d_i, \theta_j) = 1$ for $i \neq j$. Now consider two problems with the *same* parameter and decision spaces but *different* sample spaces, X_1 and X_2, and hence different densities $p_1(x_1 \mid \theta_i)$ and $p_2(x_2 \mid \theta_i)$. The analyses will proceed using values (α_1, β_1) and (α_2, β_2) respectively. (In the special Neyman-Pearson form, $\alpha_1 = \alpha_2$: if minimax is used, $\alpha_1 = \beta_1$, $\alpha_2 = \beta_2$.) Now consider the situation where X_1 arises with probability ½, as does X_2. (This sounds strange but is possible with two sample sizes n_1 and n_2: the chance mechanism with probabilities ½ is ancillary.) The "natural" values of (α, β) for the mixed experiment are $\frac{1}{2}(\alpha_1 + \alpha_2)$ and $\frac{1}{2}(\beta_1 + \beta_2)$. It can happen, even when fixing α or using minimax, that the natural values are not even admissible. The only procedure that ensures admissibility is the Bayesian method that minimizes $\pi_0\alpha + \pi_1\beta$. (That this is Bayesian follows from the normal method of analysis using the average risk, equation (3).)

A simple way to see this is to think of all possible values, including inadmissible ones, of α and β in the unit square. It is necessary to express preferences amongst these. Admissibility alone rules out any pair (α, β) for which there is available another pair (α', β') with $\alpha' < \alpha$, $\beta' < \beta$. Preferences may be expressed in terms of "contours" in the (α, β)-square along which all values are equally attractive. Mixing the values with probabilities ½ as in the previous paragraph easily shows that the contours must be parallel lines, $\alpha + k\beta$ constant, for some positive k. The value $k = \pi_1/\pi_0$ gives the Bayesian interpretation.

A more detailed account of the material in the last two paragraphs will be found in Section 3 of Lindley (1972) based on an approach developed with Savage. The famous counter-example of Cox (1958) on mixing experiments illustrates the problem. Cohen (1958), Cornfield (1969) and Bartholomew (1967) are also relevant and the last has an interesting discussion. That fixing α, as Neyman and Pearson suggested, is unsatisfactory has been long appreciated. As Anderson (1987) says "the appropriate significance level should be adjusted to sample size". Unfortunately, we are not told how the adjustment is to be made. Berger and Delampady (1987) give the Bayesian answer. That $k = \pi_1/\pi_0$ above does not change with the sample size is reflected in the optimality result for SPRT that holds the limiting values of the likelihood ratio constant as sampling proceeds. Here is coherence at work: it is missing from the rest of Wald's writings.

1.6 Summary

The discussion has pointed out that although the sampling-theoretic and Bayesian views share

parameter and decision spaces in addition to the data, the former needs to include a sample space and the latter has to introduce a distribution over parameter space. Both these introductions present problems that have not been completely resolved. The additional consideration of how separate decision problems fit together, or cohere, shows that only the Bayesian attitude is coherent. Consequently the sample space is irrelevant. (This leads to the likelihood principle, a topic that will not be discussed here: see Berger and Wolpert (1984) and Basu (1988).) My thesis is that sampling-theorists have failed to consider coherence and, in consequence, have produced unsatisfactory methods.

The Bayesian view is sometimes presented as though it just consists of adding a prior to the Waldean framework. Thus, in point estimation, a Bayes estimate is merely one that minimizes the weighted risk (3). I want to argue that the Bayesian paradigm involves a very different approach from the Berkeley one. It really is a distinct paradigm and the substitution of one by the other requires a true scientific revolution in Kuhn's (1974) sense.

2. INFERENCE

2.1 The Basic Problem of Inference

To appreciate fully the distinctive nature of the Bayesian paradigm, it is necessary to consider the foundations of statistics and the nature of the practical problems that statisticians are trying to solve. "Those of us who are concerned with our job prospects and publication lists avoid carefully the conceptually difficult problems associated with the foundations of our subject" (Lavis and Milligan, 1985). Towards the end of my career, neither is an impediment, and indulgence in the foundations may perhaps be permitted; especially when they have so much to tell us and can exert such a strong influence on our mathematics and our practice.

It is convenient to begin, not with decision-making, but with inference; or what is often called inductive logic. This will lead naturally into decision aspects. Ordinary logic is inadequate to justify something we all do when we suppose tomorrow will be, in most respects, like yesterday. Ordinary logic enables us to deduce from a hypothesis H consequences x and y. It does not help us to prove H, or to deduce that consequence y may follow from x. We need inference to pass from x (yesterday) to y (tomorrow). The fundamental problem of statistical inference is to pass from one set x of observations to express opinion about another, as yet unobserved, set y. Statisticians are familiar with it in the form where x is a set of n observations $(x_1, x_2, \cdots, x_n)$ of a series and $y = x_{n+1}$, a further value of the series. An important, special case is where the $n + 1$ values are replicates (the technical term is exchangeable), and it is desired to infer the value of one of them from observation of the remaining n. A basic problem is how to accomplish this.

2.2 The Bayesian Description of Inference

The Bayesian solution is to describe the connection between x and y by a probability $p(y \mid x)$ of y, given x. (In fact, there is always present in addition to x and y background information K, and in full one should write $p(y \mid x, K)$ but the knowledge K, being ubiquitous, is usually omitted from the notation.) According to this view, all manipulations in inference are solely and entirely within the calculus of probability. The mathematics is that of probability. The interpretation of the probability is that of the degree of belief of a subject, conveniently called You, in y when You know x (and K). The interpretation is neither classical, in terms of equally likely cases, nor frequentist.

Consider the simple case just mentioned where x is $x^{(n)} = (x_1, x_2, \cdots, x_n)$, being n replicates, and $y = x_{n+1}$, a further instance. Then You require $p(x_{n+1} \mid x^{(n)})$, or equivalently $p(x^{(n+1)})/p(x^{(n)})$. To include all n, You need to specify $p(x^{(n)})$ for all n. This is difficult—try it in the simplest case where x_i is either 0 or 1. Statisticians have found a way of doing this in the case of replicates. Suppose that there is a vector θ such that, given θ, the x_i are iid. That is, $p(x^{(n)} \mid \theta) = \prod_1^n q(x_i \mid \theta)$. All that is now necessary is for You to specify θ and q, although the value of θ being unknown, it will require a probability $\pi(\theta)$. Then

$$p(x^{(n)}) = \int_\Theta \prod_1^n q(x_i \mid \theta)\pi(\theta)\, d\theta$$

and $p(x_{n+1} \mid x^{(n)})$ can be found. In other words, the introduction of an additional ingredient, the parameter θ, enormously simplifies the probability calculations. This form, with $x^{(n)}$ and θ, is that familiar to statisticians. Formulated this way, it shows that there is no essential difference between the probability $q(x \mid \theta)$ and the prior $\pi(\theta)$: both are expressions of Your uncertainty. Of course, K may contain information that can pin down q and/or π; but often this is not so. It also shows that θ is a construct, and not an ingredient in the original, practical problem. (de Finetti, 1937, has shown that if the x's are replicates in the sense of being exchangeable then this iid structure is the only possible one.)

2.3 Scientific Inference

The statistical procedure of parameterization just described is related to one that is much used in science; indeed, it may be viewed as a generalization of scientific method. Faced with a lot of data, $x^{(n)}$ (not

necessarily replicates), science typically constructs a hypothesis H to explain the data, and then devises an experiment to test H, the result x_{n+1} of which may, or may not, support H. The Bayesian description is

$$p(H \mid x^{(n+1)}) = \frac{p(x_{n+1} \mid x^{(n)}, H)p(H \mid x^{(n)})}{p(x_{n+1} \mid x^{(n)})}.$$

It is usually supposed that, given H, x_{n+1} and $x^{(n)}$ are independent (compare the use of independence above) when

$$(5) \qquad p(H \mid x^{(n+1)}) = \frac{p(x_{n+1} \mid H)p(H \mid x^{(n)})}{p(x_{n+1} \mid x^{(n)})}.$$

The experimental result x_{n+1} will support H, in the sense of increasing Your belief in H, if $p(x_{n+1} \mid H)$ exceeds $p(x_{n+1} \mid x^{(n)})$: otherwise support will be decreased. A common case is where the result x_{n+1} is implied by H, so that $p(x_{n+1} \mid H) = 1$ and the probability of H increases. This development shows how repeated observation of results predicted by a hypothesis increases Your belief in that hypothesis.

There is thus a close connection between the scientist's use of a hypothesis H and the statistician's introduction of a parameter θ. Both make the probability calculations easier by invoking independence and hence simplify the inference. The main difference is that typically $p(x_{n+1} \mid H) = 1$ (or 0) whereas $p(x_{n+1} \mid \theta)$ is not so restricted but can take all values in the unit interval. Notice the important role played by independence. Are there any statistical models that do not utilize the concept? Even stochastic process theory has its underlying "errors" which are independent. Consider, for example, the many generalizations of the simple, linear, autoregressive process $x_{n+1} = ax_n + e_{n+1}$, where the dependent x's are expressed in terms of iid errors, the e's. The concept of iid is related to the use of frequency notions in connection with probability.

2.4 Alternative Hypotheses

There is no distinction within the Bayesian paradigm between the various values of the parameter except insofar as they may be ascribed different probabilities $\pi(\theta)$. Indeed, the coherent view is essentially one of contrast between the various values of θ. This is most clearly seen in the case where the parameter space contains just two values (θ_0, θ_1) and Bayes theorem in odds form reads

$$\frac{p(\theta_0 \mid x)}{p(\theta_1 \mid x)} = \frac{p(x \mid \theta_0)\pi(\theta_0)}{p(x \mid \theta_1)\pi(\theta_1)}$$

for data x. The data affect the change of belief through the likelihood ratio, comparing the probabilities of the data on θ_0 and on θ_1. This is in contrast with a sampling-theory (or tail-area) significance test where only the null hypothesis (say θ_0) is considered by the user of the test. Of course, attempts to justify these tests have had to consider the alternatives but the user is freed from this necessity. In a paper to be discussed later, Box (1980) has exploited this one-sided nature of tests to study model fit.

The scientist, like the statistical practitioner, only uses the single hypothesis H and rarely considers alternatives. What, for example, is an alternative to Einstein's theory? How can the alternatives be avoided? The answer is illuminating and is due to Jeffreys (1939) and Huzurbazar (1955). For simplicity, consider the discrete case.

Let $x_1, x_2, \cdots$ be a sequence of results all implied by H, so that $p(x_i \mid H) = 1$ for all i. Then equation (5) reads

$$(6) \qquad p(H \mid x^{(n+1)}) = p(H \mid x^{(n)})/p(x_{n+1} \mid x^{(n)}).$$

Since $p(x_{n+1} \mid x^{(n)})$ does not exceed one, the sequence $p(H \mid x^{(n)})$ is non-decreasing and therefore tends to a limit P, $0 < P \leq 1$. (It is supposed that $p(H)$ is not 0.) Furthermore, $p(x_{n+1} \mid x^{(n)})$ must tend to 1, otherwise the left-hand side of (6) will exceed 1, which is impossible. In words, if n results implied by H have been observed, the probability of another result implied by H tends to 1 with n.

More can usefully be said. Let $A = x^{(n)}$ and $B = (x_{n+1}, \cdots, x_{n+m})$, still with $p(x_i \mid H) = 1$. Then in generalization of (6)

$$p(H \mid A, B) = p(H \mid A)/p(B \mid A)$$

and, by a similar argument, $p(B \mid A)$ tends to 1 as m and n both tend to infinity. Thus, not just one, but m implications of H approach certainty. All this is without reference to an alternative to H. Furthermore, the final result does not involve H, but only the observable x's, and is truly an answer to the basic problem of inference (Section 2.1) when logical implication of x_i by H holds. The same facility is not available to the statistician when the implication, $p(x_i \mid H) = 1$ is replaced by $p(x_i \mid \theta)$ taking any values in the unit interval.

2.5 Theories and Models

Scientists are usually concerned with theories: statisticians are interested in models. A *theory* is a statement of wide scope that applies to many situations: a *model* is confined to a measurement process. The distinction is similar to that between strategy, with its overall view, and tactics, with its narrower concern. A theory predicts the value of θ in this situation, and ϕ in that. To measure θ and ϕ, two parametric models

are needed, each specific to the measurement process. Often a theory and a model occur together: the hypothesis H implying a value θ_0 for a parameter θ, which can be measured with error producing a result x. Testing H is equivalent to testing $\theta = \theta_0$ but with alternatives $\theta \neq \theta_0$ suggested by the model, not by the theory. Sturrock (1973) gives an example wherein θ is the power output of a nebula and x is a measurement of that power.

Notice the role of a model. It is a device that assists You in Your probability calculations, as we saw in Section 2.2. Typically, the model extends the conversation from the data to include extra quantities, parameters, that impose a simpler structure, usually through independence, on the problem.

2.6 Large and Small Worlds

An important part of statistics is concerned with the testing of a model. The chi-square, goodness-of-fit test is the most famous example. Box (1980), in a most thought-provoking paper, has suggested that the Bayesian paradigm is appropriate for analyses within a model but is inadequate for testing a model. In the latter situation, there are no immediate alternatives and Box proposes using sampling-theory ideas, in particular tail-area significance tests. I argue now that model-testing can be accommodated within the Bayesian viewpoint.

When You, a scientist or statistician, consider an inference You do not take into account everything but concentrate on what appears to You to be the relevant issues. You build for Yourself a *small world* including some data, excluding others, and within that small world construct a model for Your beliefs. Now such a world is part of a *larger world* and what You say about the small world may appear incoherent when viewed in the larger perspective. Shafer (1986) gives a good example. Consequently it is appropriate to test the small world and consider what might happen were You to enlarge it. This is possible by including more data or with extra parameters. A coherent test of that small world against a possible, larger one is now possible within the coherent view, and no sampling-theoretic considerations are necessary. Of course, You have to contemplate how the small world is to be enlarged and what the alternatives to it are. But this is surely essential. We may be surprised if $p(x \mid \theta)$ is tiny within the small world but unless there is a θ' with $p(x \mid \theta')$ bigger within the larger world, then θ must remain a plausible value and the surprise must be accepted. This helps to explain how a theory H can continue to be used even when it has been rejected. What may have happened is that H predicts $\theta = \theta_0$ and experiment establishes that $\theta \neq \theta_0$ by a significance test. Yet there is no H' to explain any value of θ other than θ_0.

2.7 Summary

An essential activity of all life is to make judgments about as yet unobserved data y on the basis of observed data x. This is the problem of inference or inductive logic. The Bayesian paradigm requires that this be done solely and entirely within the calculus of probability; in particular, the above judgment is $p(y \mid x)$. The calculation of such probabilities is substantially assisted by the consideration of theories including hypotheses H, and models incorporating a parameter θ. Independence, conditional on H or on θ, appears to be basic to the calculations. Statistical practice ought therefore to start from the data, x and y, and regard the analysis, involving theories and models, as means of evaluating the probabilities.

An important question remains: why use probability? What is the justification for the Bayesian paradigm? It is to this question that we now turn, and to the relationship between inductive logic and decision-making.

3. PROBABILITY

3.1 Why Probability?

The Bayesian paradigm requires the uncertainty of y, given x, to be described by probability, $p(y \mid x)$. With a parametric model, $p(\theta \mid x)$ is held to be the appropriate description. The Berkeley position is that statements about parameters can be made by various methods, one of which is confidence. It is important to notice that a confidence interval for θ is *not* a probability statement about θ: it is one about the interval and is derived from those about x, $p(x \mid \theta)$. Consequently, confidence is an alternative measure of uncertainty. Similarly, tolerance intervals can substitute for the Bayesian's $p(y \mid x)$. Other alternatives exist: belief functions (Shafer, 1976); fuzzy logic (Zadeh, 1983). Why use probability? To phrase the question differently: You are uncertain about y, given x; why measure this uncertainty by numbers that obey the laws of the calculus of probability, rather than other laws like those of fuzzy logic? This rephrased form of the question is important because, as we shall see, it is the laws of combination of uncertainty statements that are the key issues. We have had a partial answer to our question when we showed in Section 1.5 that the coherent selection of a unique solution from a complete class required a probabilistic description of the parameters. But Wald assumed a probabilistic description of the data: why? He also introduced loss and minimized its expected value: why? All these questions have simple answers.

3.2 The Axiomatic Approach

The answer proceeds along the following lines. Consider simple situations of uncertainty and see whether

in them there are not properties that the measure of uncertainty ought to possess. For example, transitivity: if A is more uncertain than B, and B than C, then A is more uncertain than C. These properties can then be expressed formally and used as axioms for a mathematical system that may be used to prove theorems about the measures. A basic theorem that comes out of such an analysis is that the measure of uncertainty, or a transform thereof, obeys the rules of probability. This needs a qualification: it obviously depends on the choice of axioms. Reasonable axioms lead to the statement just made. I know of no axiom system that leads to confidence measures. Other axiom systems lead to variants of the probability approach: for example, to upper and lower probabilities (Smith, 1961). These are defective for me because they do not incorporate the notion of a *unique* recommendation. Like Wald, they only produce a *class* of procedures.

The "inevitability of probability" is strengthened by the fact that three distinct axiom systems lead to that result. The first is firmly based on decision theory and is due to Ramsey (1931) and independently to Savage (1954). The second uses scoring rules, originating with de Finetti (1974/5) who also, in common with others, used a method based on betting. A third follows the usual mensuration procedure of comparison with a standard (Pratt, Raiffa and Schlaifer, 1964). There is an admirable survey by Shafer (1986; with discussion) that concentrates on Savage's method. Since the consequences of the axioms are so important, it is sensible that the axioms be subjected to the most careful scrutiny; this has been done by Fishburn (1986), again with discussion. Kolmogorov's (1933) enormous contribution was to provide axioms for probability. This work pushes the axioms further back and provides Kolmogorov's as theorems.

3.3 Paradigms

It is a side issue that these approaches, particularly that of Savage, with their axiomatic approach and rigorous development, are very much in the spirit of modern mathematics. One would therefore have expected them to have appealed to mathematical statisticians, even if the practitioners ignored them. Yet they do not, and the mathematical members of the statistical community remain largely within the Waldean scheme. (As will be seen later, it is the practitioners, or at least some of them, who notice the Bayesian aroma.)

The resolution of this paradox seems to lie in the fact that most mathematical statisticians are primarily technicians. Just as scientists, according to Kuhn (1974), work within a paradigm, rarely questioning it, so many mathematicians work within a system without concerning themselves with the reasoning behind it. This is of little importance in pure mathematics, but in an applied subject like statistics it can be unfortunate. Science, again following Kuhn, has revolutions in which the paradigm changes. The Bayesian view, firmly based on probability, is a new paradigm. All I ask of readers of these words is that they take time off from their technicalities and think about these two paradigms of statistics: Berkeley and Bayes, sampling-theoretic and coherent.

3.4 The Coherent Paradigm

To return to the main issue of Section 3.2: to fit together, or cohere, the uncertainty statements need to be probabilistic. Hence inductive logic should be expressed through probability, and inference about uncertain y from observed x is made by $p(y \mid x)$. All calculations, in order to achieve coherence, must be within the probability calculus. Your problem is to evaluate this probability using only that calculus. In statistical contexts this is almost always done through a model using a parameter θ. (Notice, in the sense in which that term is being used here, this embraces nonparametric statistics; θ indexing, for example, all distributions on the real line.) The principal effect of the parameter is to introduce independence into the system, as we saw in Section 2.2.

It is usual in statistical practice to make statements about the parameter rather than the observation y. This is sensible because such a parametric statement is available for *any* future observation z whose uncertainty depends on θ. With $p(\theta \mid x)$ from Bayes theorem, and $p(z \mid \theta)$, we have

$$p(z \mid x) = \int_\Theta p(z \mid \theta) p(\theta \mid x)\, d\theta,$$

assuming the independence of x and z given θ. This attention to θ, rather than y, is an example of good practice within the Berkeley paradigm persisting in the alternative one.

There is a disadvantage in confining uncertainty statements to parameters, and that is that they can rarely be checked against reality, simply because the value of θ is seldom ever known. It is rare to learn the exact value of the mean of that normal distribution. Whereas future observations, y or z, will be experienced and so a check on their probabilities is possible; for example, by scoring rules. A statistician who consistently gave low probabilities for observed y's has not done as good a job as one whose probabilities were higher. de Finetti has often expressed the view that probabilities should deal with observables and not abstractions like parameters. Geisser, in many papers, for example (1985), has emphasized the predictive aspect of our subject.

We now explore some consequences of the Bayesian paradigm with its emphasis on coherence. Coherence is merely the fitting together of uncertainty

judgments, and is achieved by the probability calculus. Science proceeds by fitting together a series of experimental and observational results, combining this result with that, and only ignoring a result with justification. It is surprising that statistics contains so little material within the sampling-theoretic framework on how to do this combination. How do you combine two confidence intervals for the same parameter, or the results of two significance tests for the same hypothesis? In the new paradigm, this combination is basic.

3.5 Point Estimation

In Waldean theory, point estimation is the case when D and Θ, the decision and parameter spaces, coincide. (We ignore nuisance parameters for the moment.) Required is a decision d about the value of θ. This is accomplished by a decision function $\delta(x)$ that prescribes what d to take when x is observed, as a function of x. It is a point estimator. Bayesian calculations need operate only with the observed x and do not need to use $\delta(\cdot)$: compare equations (1) and (2). Bayesian theory does not use the notion of an estimator. Within that theory, it is enough to quote the density $p(\theta \mid x)$ of θ. There is no need to introduce a decision space or a loss function.

It is fashionable today to speak of a Bayes estimate, meaning one that has been derived from Waldean theory using a loss function, usually squared error for real θ, and utilising a prior, purely as a technical device, to be discarded when studying the properties of the resulting estimator $\delta(x)$. The terminology is unfortunate since the Bayesian contribution is so minimal. There is only one Bayes "estimate," namely the full distribution $p(\theta \mid x)$. You may wish to summarize this through its mean and variance, but this is a slight problem compared with the problem of determining the "best" estimate. In fact, in the coherent view, the whole problem of finding a good point estimate disappears. There is only one estimate: $p(\theta \mid x)$. Thus a whole branch of statistics disappears. There is nothing new in this phenomenon of a whole branch of mathematics vanishing. The history of mathematics is full of topics that are today virtually ignored—who discusses quarternions? Aside from the question of best, there is much that is useful in point estimation theory and can be used in the Bayesian paradigm: sufficiency, for example.

The following, simple example illustrates difficulties with point estimation. Suppose $x = (x_1, x_2, \ldots, x_n)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, both vectors of n real numbers. Suppose each θ_i is $N(0, 1)$ independently of the others. Finally, suppose, given θ, x_i is $N(\theta_i, 1)$ again independent of the others. (It is a simple model for measurements x_i made on treatments θ_i drawn from a pool of treatments.) Two easy results are that x_i is $N(0, 2)$ marginally, and that, given x, θ_i is $N(\frac{1}{2} x_i, \frac{1}{2})$ by Bayes theorem. This latter result has suggested the use of the mean $\frac{1}{2} x_i = \hat{\theta}_i$, say, as a point estimate of θ_i. But then it has been noticed that $n^{-1} \sum \hat{\theta}_i^2 = (4n)^{-1} \sum x_i^2$ tends to $\frac{1}{2}$ as n tends to infinity, whereas $n^{-1} \sum \theta_i^2$ tends to 1. In other words, the estimates are not as dispersed as the true values.

The difficulty arises because of the use of point-estimation ideas. Bayesian concepts invite consideration of the probability density of $n^{-1} \sum \theta_i^2$, given the data x. Its mean is $n^{-1} \sum (\frac{1}{4} x_i^2 + \frac{1}{2})$ which tends to 1 as n tends to infinity and the perceived difficulty disappears. I know of no situation in which operations strictly in accord with the calculus of probability give other than sensible conclusions. Whereas ideas outside that calculus may present anomalies, as here.

If nuisance parameters, λ, are present they can easily be eliminated by integration. From the joint distribution $p(\theta, \lambda \mid x)$, You obtain

$$p(\theta \mid x) = \int p(\theta, \lambda \mid x)\, d\lambda.$$

A stumbling block in some aspects of sampling-theory statistics and the likelihood approach is thereby removed. The actual integration, analytic or numeric, may present difficulties.

It is not necessary to say anything here about hypothesis testing within the coherent picture since the topic has been admirably discussed by Berger and Delampady (1987).

3.6 Counter-Examples

To this audience, something needs to be said about the counter-examples to Bayesian ideas that have appeared in the literature. A recent, well-known paper is by Diaconis and Freedman (1986). Therein examples of Bayes estimates that are inconsistent are provided in situations where there are consistent estimates. These cannot be dismissed on the grounds that point estimation is misguided (Section 3.5), because the disturbing property can be rephrased in terms of the posterior distribution converging to something other than the true value. The examples concern mathematically complicated situations, far removed from the basic axioms, and the resolution of the paradox they create rests with the mathematics, especially with the mathematics of the infinite. We discuss a simple example of the role of infinity in mathematics.

Consider the concept of summation. With a finite collection of numbers a_i there is no real difficulty. $\sum_1^n a_i$ is easily found and, for example, the numbers may be taken in any order. Difficulties arise when the set is enumerably infinite. Here $\sum_1^\infty a_i$ does not always have a meaning and the order of summation may matter. It took mathematicians a while to define and

understand absolute convergence, when the problems largely disappear. More difficulties arise when non-enumerable summations in the form of integrals $\int a(t)\,dt$ are contemplated. The history of mathematics is full of different types of integral that try to ape the properties of finite sums. Eventually the Lebesgue integral emerges as a reasonable workable form and Fubini's theorem tells us that, for positive, integrable functions, the order of integration is irrelevant. My point is that in introducing the useful concepts of infinity and continuity, mathematicians try to carry forward the ideas present in finite, discrete cases; the anomalies, like divergent series, are ordinarily discarded. We should not be surprised at mathematical ingenuity producing inconsistent Bayes estimates, any more than we are at series which behave anomalously. The mathematical task is to provide conditions under which the properties of the finite reality extend to the infinite and continuous abstraction.

(Since almost all sampling-theoretic ideas are incoherent, it is easy to produce counter-examples to them which are much simpler and direct than those just discussed. I provided a collection in Section 3 of Lindley (1972). If a comparison of alternatives (Section 2.4 here) is the proper way to make judgments, the Bayesian argument is the clear winner in respect of counter-examples.)

3.7 Summary

A justification for the Bayesian paradigm rests on the development from basic properties of Your appreciation of uncertainty used as axioms. (An alternative justification, to be considered in Section 5.5, is that it works.) The central idea is the concept of coherence between uncertainty judgments. If you make some probability statements, then others are implied by the calculus of probability, and effectively You have made those as well (de Finetti, 1974/5, Section 3.9). The result is a paradigm which is markedly distinct from the sampling-theoretic one that is currently popular. Whilst several ideas in the latter can be carried over to the coherent view, others do not transfer and are seen to be incoherent.

The discussion has been confined to inference: Your understanding of the world. But, as Karl Marx said, "The point is not merely to understand the world, but to change it." Change implies action and decision. We now describe how the Bayesian view easily extends to decision-making.

4. UTILITY

4.1 Maximization of Expected Utility

The formulation has so far included an observed x and an unobserved y in a space Y. These are connected by $p(y \mid x)$. To include decision-making, a set D of decisions, or acts, d is added. Choice of an act will affect y, so You now have $p(y \mid x, d)$. It may be, and usually is, useful to include parameters, but this is unnecessary in the general overview now being given. The pair (d, y) constitutes a consequence c, and it is basic that You prefer some consequences to others. The procedure is as with uncertainty; basic properties of the preference pattern are used as axioms from which theorems are developed. I now outline the development of the principle of maximization of expected utility (MEU) because, although it is in the literature (DeGroot, 1970), it will demonstrate two important points that the statistical community may not have appreciated.

Suppose that there is an overall best, $\bar{c}$, and an overall worst, $\underline{c}$, consequence amongst all the (d, y) You are contemplating. (This will be true if D and Y are finite. It is a mathematical problem, of the type discussed in Section 3.6, to extend the argument to infinities.) Consider any consequence c and the choice between two outcomes,

(a) c for sure, and
(b) $\bar{c}$ with probability u, and $\underline{c}$ with probability $1 - u$.

If the preference assumption is extended to embrace mixtures like (b), there must exist a unique u such that You are indifferent between (a) and (b). Write this value $u(c)$ or $u(d, y)$. It is called the *utility* of $c = (d, y)$.

If d is selected, y, and hence $c = (d, y)$, will have probability $p(y \mid d, x)$. But c may be replaced by $\bar{c}$ or $\underline{c}$ with probabilities $u(c)$ and $1 - u(c)$ by the above indifference. By a basic rule of probability, the result of selecting d is that $\bar{c}$ will equivalently be obtained with probability

$$\bar{u}(d, x) = \sum_y u(d, y)p(y \mid d, x). \tag{7}$$

($\underline{c}$ will be obtained with probability $1 - \bar{u}(d, x)$.) $\bar{u}(d, x)$ is called the *expected* utility of d (given x), because it is a genuine expectation with respect to the density $p(y \mid d, x)$. You naturally wish to maximize Your probability of $\bar{c}$, as against $\underline{c}$, so You should choose d to maximize Your expected utility. The demonstration is complete.

4.2 Utility as Probability

The first point that emerges from this argument is that utility is well-defined as a probability that equates a gamble (b) with a sure consequence (a). It has never been clear what Wald's loss was. It is usually supposed by adherents of utility theory that $L(d, y) = \max_d u(d, y) - u(d, y)$, the difference between the best decision for y and that for the decision selected.

If so, MEU is equivalent to minimization of expected loss (over y).

The second point follows from this description of utility in terms of probability, namely that utility conforms to the inviolate rules of the probability calculus. Hence $\bar{u}(d, x)$ can be calculated as in (7) and is immediately seen to be an expectation. Thus the expectation arises naturally and is seen to be the *only* feature needed in order to choose the best d. It was never clear why only expected loss should be considered.

4.3 The Complete Bayesian Paradigm

The coherent approach is now complete, incorporating decision-making as well as inference. With data x (and background information K) You contemplate another data set y in Y and an action d in D that will influence y. Your uncertainty of y is described by $p(y \mid d, x)$. Your preferences among consequences are described by $u(d, y)$. The best decision is MEU

$$\max_d \sum_y u(d, y)p(y \mid d, x).$$

The calculations of probabilities and utilities are often helped by the introduction of a parameter θ. This ordinarily has the effect of making x and y independent given θ and d (Section 2.2). It also typically happens that θ is unaffected by the choice of d so that $p(\theta \mid d, x) = p(\theta \mid x)$. If both these obtain then

$$\begin{aligned}\bar{u}(d, x) &= \sum_y u(d, y)p(y \mid d, x) \\ &= \sum_y u(d, y) \sum_\theta p(y \mid d, \theta)p(\theta \mid x)\end{aligned}$$

on extending the conversation to include θ and using both independence properties. So, on interchanging the orders of summation,

$$\bar{u}(d, x) = \sum_\theta u^*(d, \theta)p(\theta \mid x) \tag{8}$$

where

$$u^*(d, \theta) = \sum_y u(d, y)p(y \mid d, \theta).$$

This means that $u^*(d, \theta)$ can replace $u(d, y)$ and y ignored.

In this parametric form, *inference* is the calculation of $p(\theta \mid x)$. *Decision-making* is the combination of this with the utility $u^*(d, \theta)$ and MEU (equation 8). Notice how the inferential process is usefully separated from the decision activity but is available for *any* decision that involves θ. Some inferential procedures, like confidence intervals, are inadequate because they do not fit into any decision framework.

Notice how constructive the paradigm is. It is like a recipe. You only have to follow the rules. What do You know? x and K. What is uncertain? y (or θ). What are the possible decisions? d. The recipe is to calculate $p(y \mid x)$ and $u(d, y)$ (or $p(\theta \mid x)$ and $u^*(d, \theta)$) and choose that decision that maximizes the expectation of the latter with respect to the former. In the coherent system it is perfectly clear what has to be done. The difficulties are the evaluation of some of the probabilities and utilities, and the calculation of others. The latter is a problem within the probability calculus and has been much studied. The former has not been adequately treated, and we will return to it in Section 6.

4.4 Expectation as the Sole Criterion

Objections have been raised to the sole use of expectation as a criterion of choice. People have felt the need to include, for example, the variance as well. I think this arises because of a misunderstanding about the nature of utility. Statisticians have paid scant attention to utility (or loss) and the total literature is not vast.

The most famous objection is due to Allais (1987) who argues that the utility of a monetary prize may be affected by the probability of that prize: thus, an unexpected 1,000 dollars may have more utility than an anticipated 1,000. He expresses this by saying that the probabilities and utilities may not be independent. As a description of people's behaviour, this is undoubtably correct, but it does not fault the use of expectation. Utility is attached to a consequence (d, y), or (d, θ), and in evaluating that consequence You are free to take into account any aspect of it that You wish. In particular, You may wish to include the surprise that will delight You if d results in the unexpected y. What is happening is that You are not just considering money but other aspects of the situation as well. Once relevant aspects are included the difficulty disappears.

Another point about utility is that every utility is an expected utility. Our discussion has been in a small world (Section 2.6); utility in that small world is an expectation over a larger world that contains it. The point just described can be expressed as saying that money is too small a world.

4.5 Probability as the Sole Measure of Uncertainty

Just as expected utility has been criticized for bearing too heavy a burden for choice, so probability has been attacked for being an inadequate description of uncertainty. For example, faced with an uncertain event E about which You know very little, You may tentatively assign it a probability of ½. Whereas You will confidently assign ½ to the chance of a reputable coin falling heads. The first ½, it is argued, is different from the second, and probability fails to recognize this. I argue that, on the contrary, probability theory

will recognize this difference if it is relevant to the decision at hand. Here is an example.

A bag contains $2N$ balls, where N is large, some of which are white, the rest are black. A ball is to be drawn at random and if it is white You will receive 100 dollars, otherwise nothing. There are two variants:

(a) You know N of the balls are white and N black.
(b) You are uncertain about how many balls are white but think that all values 0, 1, 2, ..., $2N$ are equally probable; the remainder being black.

In both variants Your probability of a single, drawn ball being white is ½, yet most people prefer (a) to (b) because there is more uncertainty in (b) than in (a). But now change the situation so that 2 balls are to be drawn and the 100 dollars obtained only if the colours of them match. The probabilities of winning are ½ in (a) and ⅔ in (b). Hence the calculus of probability incorporates the additional uncertainty of (b) over (a) if that uncertainty is relevant, as it is when 2 balls are drawn, though not with a single drawing. An alternative description is to think of the single drawing as a small world embedded in the larger world of 2 balls being drawn.

4.6 Summary

The Bayesian paradigm is a complete recipe for appreciation of the world by You, and for Your action within it. The central concept is probability as the sole measure of uncertainty and as a means of expressing Your preferences through (expected) utility. Properly appreciated, both measures are adequate for inference and decision-making. The coherent view stands in marked contrast to the sampling-theoretic one, and we now study this contrast in more detail.

5. A NEW PARADIGM

5.1 Prior Probability

The modern version of the Bayesian paradigm has been around now for at least a third of a century, since Savage's (1954) book, and as an operational tool for almost half a century (Jeffreys, 1939). Although the statistical literature contains more papers than it used to in the Bayesian vein, the paradigm has not achieved the success it theoretically deserves. It is therefore worthwhile to consider why this is so, and why we have not had Kuhn's revolution. The usual response is: because of the difficulty of determining the "prior," $\pi(\theta)$. So let us begin with this.

In the Bayesian view all probabilities are alike, they are all expressions of Your beliefs. There is no qualitative difference between $p(x \mid \theta)$ and $\pi(\theta)$ for data x and parameter θ. They are both probabilities based on past experience, K. Sometimes K contains information that clearly indicates $p(x \mid \theta)$ is normal or Poisson. Less often it tells us about $\pi(\theta)$, as in sampling inspection or empirical Bayes situations. But there are many occasions where such information is lacking, and Berkeley supposes normality for convenience. It is honest and recognizes this, for example, in the development of robust procedures or nonparametric methods. Why cannot Berkeley do the same with $\pi(\theta)$?

Another reason for imagined difficulties with $\pi(\theta)$ is the habit that has grown up of thinking of it as a description of a Greek letter, rather than of a reality. The data x are real, θ, as explained in Section 2.2, is less so but nevertheless usually corresponds to something You can think about. Whilst it is hard, and ridiculous, to think about a prior for θ, it is easier, and sensible, to consider Your (or Your client's) opinion of the likely effect of the treatment—where effect and treatment are well-defined—because θ now has a tangible and relevant interpretation. So confusion over the prior is often due to an unnecessary level of abstraction on the part of the statistician. As de Finetti once said to me: "Stop thinking about Greek letters."

That the prior $\pi(\theta)$ is really the same in principle as the likelihood component $p(x \mid \theta)$ is easily appreciated from the fact that it is sometimes not clear what is in the likelihood and what is in the prior. The coherent attitude is that You have to specify the complete probability structure of the problem; You are free to do this in any way You wish. Eccentrically, You may prefer to assess $p(x, \theta)$ or even $p(x)$ and $p(\theta \mid x)$. A simple example of this ambiguity is provided by contrasting models I and II analyses of variance. In a one-way layout with t treatments of means $(\theta_1, \theta_2, \ldots, \theta_t) = \theta$, You may specify $p(x \mid \theta)$ and $\pi(\theta)$. This would be model I. But You may think of the θ_i as exchangeable and iid $N(\phi, \sigma^2)$ for suitable ϕ and σ^2. This is model II for the investigation of the component of variance σ^2. Is then this last density part of the prior or of the likelihood? In the coherent view, the two models are identical. The confusion between likelihood and prior has been discussed by Bayarri, DeGroot and Kadane (1988).

5.2 Linear Models

One reason for the lack of success of the Bayesian approach amongst statisticians is their failure to recognize it as a separate paradigm, distinct from their own, and merely to think of it as another branch of statistics, like linear models. Some years ago there was a conference on linear models with non-orthogonal designs. Within the sampling-theory school there are real difficulties of analysis once orthogonality is absent. No Bayesian paper was in the list that I saw, but the coherent view works equally well for all

designs. The only attractive and distinguishing feature of orthogonality is the simplicity it introduces into the calculations—essentially the matrices are easy to invert. In the Bayesian view treatment effect θ can be isolated from the nuisance parameters ϕ by taking the joint distribution $p(\theta, \phi \mid x)$ and determining its marginal by integration (Section 3.5) to obtain $p(\theta \mid x)$. With orthogonality, θ and ϕ may be independent and the integration is trivial. Many of the difficulties the Berkeley school experiences are of their own making, and they fail to recognize that the Bayesian view is so different that the difficulties disappear. Of course, I am not contending that there are no problems with the coherent attitude. There are; the computations are by no means easy. My point is that the solution is there in more than just an outline.

5.3 Multiple Comparisons

A striking example of difficulties caused by a faulty paradigm, disappearing under another, is provided by multiple comparison techniques. (Point estimation was discussed in Section 3.5.) With t treatments, there are t effect-means, $\theta_1, \theta_2, \ldots, \theta_t$, estimated by $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_t$, say. The difficulty is that if the apparently best treatment, the largest $\hat{\theta}_i$, is to be compared with the apparently worst, the smallest $\hat{\theta}_i$; the comparison cannot be carried out in the usual way, like a t-test, because of the selection of best and worst. For example, if the number of treatments is large, the difference will almost certainly be significant. A considerable literature has developed. Berry (1988) provides a Bayesian treatment.

In the Bayesian view the difficulty immediately disappears. The joint posterior distribution is $p(\theta_1, \theta_2, \ldots, \theta_t \mid x)$ and the marginal for *any* pair, $p(\theta_i, \theta_j \mid x)$, can be evaluated by integrating out the remaining parameters (Section 3.5). It matters not that i and j index the apparently largest and smallest; the integration is always valid as a rule of the probability calculus. When confronted with this analysis, students often feel it must be fallacious because $p(\theta_i \mid x)$ concentrates about $\hat{\theta}_i$, the maximum likelihood estimate. But in fact, it does not. A reasonable prior for the treatment effects nearly always leads to a shrinkage phenomenon which affects the extreme values the most. (Shrinkage, and similar phenomena like ridge regression, are most easily understood within the coherent view, and are only satisfactory when the prior that leads to them is sensible.)

5.4 Normality

It was argued in Section 5.1 that there is often little reason for choosing a specific form for $p(x \mid \theta)$. In practice, normality is often chosen for ease of calculation and the whole battery of excellent procedures that stem from it: least squares, analysis of variance, etc. The Bayesian view equally benefits from normality but, unlike Berkeley, does not demand it. At the cost of much more substantial amounts of calculation, it is perfectly possible to analyse, for example, a Latin square, with distributions having longer tails than the normal, or even being skew. All that is needed is a family of distributions with a few parameters to keep the dimensionality of the parameter space to a manageable amount. An advantage of longer tails is that outliers will be more easily accommodated in the analysis. There is no need for jackknives. The operations of the calculus will automatically damp down the influence of extreme observations. For example, with a t-distribution, as the largest member of a sample increases, the posterior distribution of the mean will shift downwards, showing the decreasing influence of that extreme observation as it becomes more extreme.

The effect of non-normality is clearly seen in the generalized linear model (McCullagh and Nelder, 1983). This model ingeniously generalizes the normal case to other members of the exponential family and provides a sensible link between the dependent and regressor variables. But deprived of the logic of least squares, it has to argue separately, employing likelihood ideas and adhockeries. A Bayesian view would have embraced the model but provided a coherent analysis, free of adhockeries, without the need for special techniques. The existing analyses, incorporated in GLIM, suffer from the same defects as least squares, in failing to provide admissible solutions, and ignoring prior knowedge, which is often important in multi-parameter problems.

5.5 Practical Statistics

The support for the Bayesian view in these lectures has been through coherence. There is another argument in its favour, the pragmatic one that it works. It is this practical test of usefulness that will eventually establish the paradigm. It is most likely through applications that the change will take place and unless departments of statistics embrace it, they will fade away because they will not be providing the better service to practitioners. The Bayesian advance is most noticeable in those branches of science that were not greatly altered by the Fisherian revolution in statistics between the wars. Agriculture, and other biological fields, retain their Fisherian traditions. Though even here there are changes. It has recently been shown by Robinson (1987) that Henderson's (1975) important work on components of variance applied to animal breeding is most easily understood within the coherent view.

By contrast, electrical engineering is dominated by the Kalman filter, and generalizations thereof, which

are completely Bayesian. My personal judgment is that the filter is the most important Bayesian advance of recent years. Essentially it is the familiar, linear model, $E(x) = A\theta$, for data x, with expectation linearly dependent, in a known way through a design matrix A, on a parameter θ. Added to this is a distribution for θ having a similar, linear structure, $E(\theta) = B\phi$ in terms of hyperparameters ϕ. As a result of the latter, the distribution of θ is not centred around the least-squares values, but usually around values which are "shrunk" towards a common value. Ridge regression is a familiar example. The "shrunken" estimates (to use the statistical nomenclature) are more efficient, in the sampling-theoretic sense. Despite its proven worth, the filter has yet to enter into the familiar statistical packages. A good exposition is by Meinhold and Singpurwalla (1983). Important generalizations are provided by West, Harrison and Migon (1985).

Other fields have long been Bayesian, in two instance since the twenties. Insurance evaluations combine past experience with data experience to produce a credibility formula (Jewell, 1974). Educational testing combines general experience of a test, expressed through its reliability, with the test result for a candidate (Lord and Novick, 1968). Both these are special cases of the Kalman filter.

Decision analysis is in a strange state of confusion. There are some statisticians, like Durbin (1987), who admit the Bayesian argument when action is needed, reserving the Berkeley position for inference. To these, I reply, what is the use of an inference if it is not available for some, perhaps unstated, decision? This is Ramsey's (1931) original riposte. Decision analysts sometimes distinguish between two classes of problem: decision-making under uncertainty (where there is supposedly no information for a distribution on parameter space) and under risk (where the distribution is known). These correspond to the two attitudes to statistics. The former fails to appreciate the coherence aspect.

An interesting field for the Bayesian approach is medicine. Much diagnostic work is within the paradigm but clinical trials, where frequency concepts have been used for many years, resists the challenge despite the works of Anscombe (1963), Bather (1985), and others. It would be interesting to know how much money has been wasted on inappropriate, incoherent analyses of clinical trials.

Law is another field in which the two views contend. Finkelstein (1978) frequents, whilst the forensic scientists believe (Evett, 1984). The role of the likelihood ratio, comparing the probabilities of the evidence, first on the presumption of guilt, second on that of innocence, is of paramount importance in interpreting evidence.

5.6 Summary

The failure of the Bayesian paradigm to enter mainstream statistics is partly due to statisticians not recognizing it as a separate and distinct paradigm with its own way of appreciating statistical problems. The paradigm justifies some procedures, like shrinkage, but dismisses others, like multiple comparisons. It is in applications that the view has made progress. Elsewhere (Lindley, 1984) I have argued that there is a bright future for statistics with the Bayesian formulation. For uncertainty and decision-making are pervasive and statisticians understand the basic tool, probability. Mosteller (1988) has similarly argued for the broad view of statistics, though without mentioning coherence. Much uncertainty is not frequentist and the Berkeley view is unsuitable.

The reason most usually given for not adopting these notions is the difficulty of assessing the prior. That there is a difficulty is undoubtably true, but it applies to the likelihood as well (Section 5.1). It is the difficulty of measuring any probability. We therefore turn to a consideration of this topic.

6. PROBABILITY ASSESSMENT

6.1 The Measurement of Probability

Consider the problem of making a map of the surface of the earth. The normative theory underlying this is Euclidean geometry. This is a system that starts with certain axioms and from which theorems are developed: for example, that the sum of the angles of a planar triangle is 180 degrees. Euclidean theory alone is not enough to make the map. It is necessary to have a method of measuring the distances and angles on the earth's surface. These measurements are disturbed by error from the "true" values and the sum of the angles of a measured triangle may not add up to exactly 180 degrees. Euclid's theory is 2,000 years old; good measurement dates from the seventeenth century, as is evidenced by Columbus's wrong evaluation of a degree of longtitude that led him to confuse America and China. Three ingredients were needed before accurate mapping became possible,

(a) the invention of triangulation,
(b) the construction of theodolites, and
(c) a calculus of errors incorporating least squares.

My thesis is that mapping the earth's surface is analogous to assessing Your uncertainty about the world. In the latter case, the normative system is not Euclid's but the calculus of probability. (It is not unreasonable to describe Savage as the Euclid of statistics, because he was the first to spell out the axioms and provide rigorous proofs.) Your measurements of probability will be subject to error so that, for example,

the uncertainties of a partition may fail to add to one, analogous to the surveyor's discrepancies in the angles of a triangle. At the moment Your probability assessments may be as rough and ready as Columbus's were of longitude. My purpose is to study the measurement of probabilities and make some tentative suggestions of how a theory might be developed. This is essentially the third ingredient above, (c). We have nothing corresponding to (b), theodolites, but the equivalents of triangulation, (a), do exist.

Suppose that You wish to determine Your probability of an event A (under conditions K which will be fixed and omitted from the notation). One way to do this is to extend the conversation to include a second, related event B:

$$p(A) = p(A \mid B)p(B) + p(A \mid \bar{B})p(\bar{B}).$$

Then $p(A \mid B)$, $p(A \mid \bar{B})$ and $p(B)$ can be assessed (assuming You use $p(\bar{B}) = 1 - p(B)$) and combined by means of this formula to provide $p(A)$. Is this *indirect* determination of $p(A)$, that includes B, in some sense better than the direct evaluation of $p(A)$? O'Hagan (1988) has emphasized the role of indirect methods in probability. Another analogue of triangulation is Bayes theorem for data x and parameter θ:

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta).$$

How does the direct evaluation of the posterior (to x) compare with the assessments of the likelihood and the prior and their combination in Bayes formula? We would look askance at someone who contemplated the posterior without going through the ritual of likelihood and prior. Is this attitude soundly based? If You can assess a prior directly, why not a posterior?

All measurement theory involves the concept of a "true" value and a measured or calculated value. The same notions are needed for Your probability of an event A. Your true probability will be written $\pi(A)$ and its direct measurement, $p(A)$. Consistently the Roman letter will be used as the measurement of the true value corresponding to the equivalent Greek letter. It is not necessary to engage in any deep discussion of what is meant by $\pi(A)$. It suffices to recognize the role it plays in the calculations. After all, at the atomic level, it becomes hard to know exactly what is meant by the length of this desk upon which I write. Notice that $\pi(A)$ is still personal to You. It is subjective. There is no suggestion of a true, impersonal probability shared by all rational persons. Also the π's are totally coherent: that is, they obey *all* the rules of the calculus of probabilities. This need not be true of the p's.

6.2 Related Work on Probability Assessment

There is an extensive literature on probability assessment. Psychologists have studied the ways in which subjects make probability judgments and have placed emphasis on their incoherent, and therefore unsatisfactory, behaviour. A comprehensive reference is the book edited by Kahneman, Slovic and Tversky (1982). Some of this work is directly relevant to the material presented in this part of these lectures because it provides experimental evidence about the types of bias and the form of the variances that might arise in practice. The psychologists' subjects are often rather naive about probability and as people become better informed, the incoherence and the errors can be expected to diminish. There is need for continual interaction between psychologists and statisticians on these matters.

Other work on probability assessment has been performed by statisticians who have pondered the question of how the most effectively to elicit probabilities from subjects. Should one ask for means and standard deviations for an uncertain quantity, or is it better to use fractiles? How, for example, can a subject assess the degrees of freedom for a t-distribution (Dickey, Dawid and Kadane 1986)? A striking practical example is contained in Kadane, Dickey, Winkler, Smith and Peters (1980). One way of assessing probabilities, suggested by de Finetti, is by means of scoring rules. These have been studied by DeGroot and Fienberg (1986) and Winkler (1986) amongst others. Any attempt to assess probabilities is liable to be involved with utility considerations, and this has been investigated by Kadane and Winkler (1988).

A referee has pointed out the similarity of purpose between the study here and experimental design. In the latter, it is required to find designs that are, in some sense, good at determining the values of quantities, usually parameters. Thus one design would be preferred to another if it was expected to produce a smaller confidence or credible interval for the parameter. In such a situation, our analysis would be concerned with the possible variability in the interval due to either the likelihood or prior being imprecisely assessed. In that sense we are concerned with robustness questions. Experimental design and the error analysis discussed here supplement one another in considering related aspects of the same problem. We shall not discuss distributions, and therefore credible intervals, but confine our attention to probabilities of events. Unlike experimental design, it will be possible to say whether such a probability is well-determined. If the proceedings in a court of law be thought of as an experiment (with the evidence as data), then we shall see that the final probability can be ill-determined.

6.3 Errors in Probability Measurement

Let p be a measurement of a true probability π. Then, again analogous to least-squares theory, it is

usual to consider two aspects of the measurement, the bias, $E(p \mid \pi) - \pi$, and the variance $V(p \mid \pi)$. These are both lightly disguised probability statements about probabilities, or measurements thereof. Again, we need not worry too much about their exact meanings anymore than the surveyor concerns himself with the probabilities underlying the biases and standard errors of his theodolites.

There is, however, an important difference between the surveyor and the probabilist. The former typically assumes that the variances are constant. This is clearly not so with probabilities, confined to the unit interval: values of π near 0 or 1 will ordinarily have smaller standard errors than those around ½. A possibility is to suppose some transform of probability, like log-odds, has constant variance. The following argument seems preferable, at least as a first stab at a theory.

ASSUMPTION 1. $V(p \mid \pi) = f(\pi)$, $E(p \mid \pi) - \pi = g(\pi)$, say, where both f and g have second derivatives. Also if p_1 measures π_1 in the presence of $\pi_2, \cdots, \pi_n$, $V(p_1 \mid \pi_1, \pi_2, \cdots, \pi_n) = V(p_1 \mid \pi_1) = f(\pi_1)$; similarly $E(p_1 \mid \pi_1, \pi_2, \cdots, \pi_n) - \pi_1 = g(\pi_1)$.

The functions f and g may change with the events, some being easier to assess than others in a way that will emerge in Theorem 1. The biases and variances are supposed not to be affected by other events. This might not always be true. Thus in density estimation, if $\pi_1, \pi_2, \cdots, \pi_n$ correspond to a fine partition of the real line, measurement by p_1 of π_1 may be affected by the closeness of π_2 to π_1 implied by the anticipated smoothness of the underlying density.

ASSUMPTION 2. $0 \leq p \leq 1$; and $\pi = 0$ (1) implies $p = 0$ (1).

The first part of this assumption trivially supposes that any measurement of a probability lies in the unit interval. The second part more substantially says that an event known by You to be false (true) will be correctly measured. There are three basic laws of probability: convex, additive and multiplicative. This assumption says that the measurements, like the true values, obey the law of convexity. The next assumption treats the addition law similarly.

ASSUMPTION 3. For exclusive events, A_1 and A_2, the measurement of $\pi(A_1 \cup A_2)$ is the sum of the measurements of $\pi(A_1)$ and $\pi(A_2)$.

A person satisfying this assumption will be called *additively coherent.* In particular, Your measurements of probabilities for a partition will add to 1. We now show that if You are additively coherent and also obey Assumptions 1 and 2, You have a special, simple error structure for Your measurements.

THEOREM. *If Assumptions* 1–3 *obtain*

$$E(p \mid \pi) = \pi \quad \textit{and} \quad V(p \mid \pi) = \kappa\pi(1 - \pi)$$

for some positive constant κ.

Let p_i be the measurement of $\pi_i = \pi(A_i)$ for exclusive A_i, $i = 1, 2, 3$. $E(p_1 \mid \pi_1, \pi_2, \pi_3)$, which is $E(p_1 \mid \pi_1)$ by assumption 1, will be written simply $E(p_1)$, and other expectations and variances will be treated similarly. Now $E(p_1 + p_2) = E(p_1) + E(p_2)$ and, by additive coherence, $p_1 + p_2$ measures $\pi_1 + \pi_2$. Hence, by Assumption 1, $g(\pi_1 + \pi_2) = g(\pi_1) + g(\pi_2)$ and $g(\pi) = c\pi$ for some c. But $g(1) = 1$ by Assumption 2, so $c = 1$ and the result concerning the expectation is proved. Similarly

$$V(p_1 + p_2) = V(p_1) + V(p_2) + 2C(p_1, p_2),$$

where $C(p_1, p_2)$ is the covariance between p_1 and p_2, so that

$$(9) \quad C(p_1, p_2) = \tfrac{1}{2}\,[f(\pi_1 + \pi_2) - f(\pi_1) - f(\pi_2)]$$

on invoking Assumptions 1 and 3. If $V(p_1 + p_2 + p_3)$ is also expressed in terms of the variances and covariances of the p_i and the latter expressed in terms of the function f by the last formula, we easily obtain, after a little rearrangement,

$$f(\pi_1 + \pi_2 + \pi_3) = f(\pi_1 + \pi_2) + f(\pi_2 + \pi_3) + f(\pi_3 + \pi_1) - f(\pi_1) - f(\pi_2) - f(\pi_3).$$

Setting all the π's to 0, the result is $f(0) = 0$. By assumption 2, $f(1) = 0$ also. Differentiation of the equation with respect first to π_1 and then to π_2 yields

$$f''(\pi_1 + \pi_2 + \pi_3) = f''(\pi_1 + \pi_2)$$

since terms without both π_1 and π_2 disappear. Hence $f''(\pi)$ is constant, $f(\pi)$ is quadratic and, with the boundary conditions already established, $f(\pi) = \kappa\pi(1 - \pi)$ as required.

COROLLARY. $C(p_1, p_2) = -\kappa\pi_1\pi_2$.

This is immediate from (9).

The variances and covariances for the p's are exactly those that would be implied by a Dirichlet distribution with $E(p_i) = \pi_i$. This may be unsatisfactory in some circumstances because it does not allow for sufficient correlation between the measurements. This lack arises from Assumption 1.

The variance form, $\kappa\pi(1 - \pi)$, would equivalently result from supposing that arc sin $p^{1/2}$ had constant variance, $\frac{1}{4}\kappa$.

Herein it will be supposed that You are additively coherent and consequently have an error structure of

the form in the theorem. This is done in order to simplify the arguments and to obtain a feel for the sort of results that might obtain. It is clearly necessary, in order to obtain a more complete picture, to investigate other moment structures. With Assumptions 2 and 3 You are obeying two of the three basic laws of probability, only the multiplication law possibly being violated. It is possible to investigate the error structure of someone who is multiplicatively, but not additively, coherent.

6.4 Extension of the Conversation

A possible method of measuring the probability $\pi(A)$ for an event A, analogous to triangulation in surveying, is to extend the conversation to a partition $(B_i; i = 1, 2, \ldots, n)$ and use the formula

$$\pi(A) = \sum_1^n \pi(A \mid B_i)\pi(B_i).$$

This indirect method is now compared with the direct assessment of $\pi(A)$. To simplify the notation, write

$$\alpha_i = \pi(A \mid B_i) \quad \text{(likelihoods)},$$

$$\beta_i = \pi(B_i) \quad \text{(prior)},$$

and

$$\alpha = \pi(A) \quad \text{(direct)}.$$

The α_i are called likelihoods because they are probabilities of A for various conditions, the B_i, whose probabilities are termed priors. (They will subsequently play these roles in the analysis of Bayes formula.) In accord with the Greek-Roman convention, a_i will be the measurement of α_i, etc. Our task is to see which is the better measurement of α, a or $\sum a_i b_i$.

It is first necessary to make assumptions about the second moments beyond those implied by Theorem 6.3.

ASSUMPTION 1. $V(a_i) = \kappa\alpha_i(1 - \alpha_i)$, $C(a_i, a_j) = 0$, $i \neq j$.

In words, each of the likelihoods is equally hard to measure since they share a common constant κ, and they are uncorrelated. This lack of correlation is perhaps a severe condition. Detailed calculations, not produced here, that follow along the lines below, show that if there is the same correlation between all pairs of likelihoods, the same general conclusions concerning the extension of the conversation persist.

ASSUMPTION 2. $V(b_i) = \lambda\beta_i(1 - \beta_i)$, $C(b_i, b_j) = -\beta_i\beta_j$, $i \neq j$.

This is similar to Assumption 1, but applied to the prior. The covariance is that implied by the partition and Theorem 6.3.

ASSUMPTION 3. $C(a_i, b_j) = 0$.

ASSUMPTION 4. $V(a) = \mu\alpha(1 - \alpha)$.

The constants κ, λ, μ describe the ease with which the likelihoods, prior and direct assessments can be made. Their relative values are important; for example, if κ and λ are large, but μ small, the extension cannot be expected to be a good, indirect method.

In evaluating the variances, we shall use the delta method, exemplified by writing differentials

$$\delta(\sum a_i b_i) = \sum a_i \delta b_i + \sum b_i \delta a_i,$$

squaring, taking expectations and regarding the expectation of squares of differentials as variances. The result here is

$$V(\sum a_i b_i) = \sum \alpha_i\alpha_j C(b_i, b_j) + \sum \beta_i\beta_j C(a_i, a_j) + \sum \alpha_i\beta_j C(a_j, b_i). \tag{10}$$

As a result of using the delta method, the values obtained will only be approximate but useful if the errors, quantified by κ, λ, μ are small. (Exact results are available with the extension of the conversation but not with the other indirect procedures considered. The differences between exact and approximate in the former case are typically slight.)

Inserting the variances and covariances of Assumptions 1–3 into (10),

$$\begin{aligned} V(\sum a_i b_i) &= \lambda \sum \alpha_i^2\beta_i(1 - \beta_i) - \lambda \sum \alpha_i\alpha_j\beta_i\beta_j \\ &\quad + \kappa \sum \beta_i^2\alpha_i(1 - \alpha_i) \\ &= \lambda \sum \alpha_i^2\beta_i - \lambda(\sum \alpha_i\beta_i)^2 \\ &\quad + \kappa \sum \beta_i^2\alpha_i(1 - \alpha_i). \end{aligned} \tag{11}$$

The direct method gives by Assumption 4,

$$V(a) = \mu\alpha(1 - \alpha) = \mu \sum \alpha_i\beta_i - \mu(\sum \alpha_i\beta_i)^2, \tag{12}$$

by coherence of the true values. Comparison of (11) and (12) depends on the values of κ, λ and μ.

THEOREM. *If* $\kappa = \lambda = \mu$

$$\begin{aligned} &V(a) - V(\sum a_i b_i) \\ &= \kappa \sum \alpha_i\beta_i - \kappa \sum \alpha_i^2\beta_i - \kappa \sum \beta_i^2\alpha_i(1 - \alpha_i) \\ &= \kappa \sum \alpha_i(1 - \alpha_i)\beta_i(1 - \beta_i) \geq 0. \end{aligned} \tag{13}$$

In words, if all measurements are equally precise, the indirect method of extension of the conversation has smaller variance than the direct assessment of the probability. Investigation of numerical cases shows that the reduction in variance can be dramatic. For example, if $\alpha_i = \alpha$, all i, and $\beta_i = n^{-1}$, the improvement, from (13), is $\kappa\alpha(1 - \alpha)(1 - n^{-1})$, whereas the direct assessment gives $\kappa\alpha(1 - \alpha)$: the reduction is by a factor of n.

The sampling-theory school often regards the likelihood as known, corresponding to $\kappa = 0$ (Assumption 1). If, in addition, $\lambda = \mu$,

$$V(a) - V(\sum a_i b_i) = \lambda \sum \alpha_i(1 - \alpha_i)\beta_i \geq 0,$$

an even greater improvement.

In view of the complexities of the calculations with Bayes formula below, it is desirable to simplify the expressions. A convenient way to do this is to think of the likelihoods as random quantities α taking values α_i with probabilities β_i. Then $\sum \alpha_i\beta_i = E(\alpha)$, the expectation of α; and $\sum \alpha_i^2\beta_i = E(\alpha^2)$. The remaining sum in (11) and in (13), $\sum \beta_i^2\alpha_i(1 - \alpha_i)$ cannot be so expressed, but the following argument shows that it is often small in comparison with the others. Let n get large, making a fine partition, in such a way that β^*, the largest of the β_i, tends to 0. Then $\sum \beta_i^2\alpha_i(1-\alpha_i) \leq \beta^* E(\alpha(1-\alpha))$ and, if the expectations remain finite, tends to 0 as n increases.

Under these circumstances, from (11)

$$V(\sum a_i b_i) = \lambda E(\alpha^2) - \lambda E(\alpha)^2, \tag{14}$$

and from (12)

$$V(a) = \mu E(\alpha) - \mu E(\alpha)^2. \tag{15}$$

It is fairly easy to see, with general μ and λ, when the indirect method (14) is preferable to the direct (15). If $\lambda = \mu$, the improvement by extension of the conversation is $\lambda E(\alpha(1 - \alpha))$. Notice that κ is absent from (14); consequently errors in the measurement of the likelihood (which κ determines) do not matter in the approximation and are therefore of less importance than errors in the prior.

6.5 The Product Law

A commonly used device for evaluating the probability $\pi(A)$ of an event A, especially when it is small, is to express A as the product of other events $A_1A_2 \cdots A_n$ and to use the product law $\pi = \pi_1\pi_2 \cdots \pi_n$, where

$$\pi_i = \pi(A_i \mid A_1, A_2, \cdots, A_{i-1}).$$

An example arises in fault-tree analysis where a fault A can only occur if faults $A_1, A_2, \cdots, A_n$ all occur. The variance of the indirect evaluation $p_1p_2 \cdots p_n$ is now compared with the direct assessment of π by p. Notice that all the p's (except that for A_1) are conditional probabilities: independence is not assumed.

ASSUMPTION 1. $V(p_i \mid \pi_i) = \kappa\pi_i(1 - \pi_i)$, $V(p \mid \pi) = \kappa\pi(1 - \pi)$ and $C(p_i, p_j) = 0$, $i \neq j$.

In words, all the probabilities have the same constant κ governing their error structure, and the component probabilities are uncorrelated.

THEOREM. *If Assumption* 1 *obtains, the indirect, product method is better, the reduction in variance being*

$$\kappa\pi(1 - \pi) - \kappa\pi^2 \sum (1 - \pi_i)/\pi_i.$$

By the delta method

$$V(p_1p_2 \cdots p_n) = \pi^2 \sum V(p_i)/\pi_i^2 = \kappa\pi^2 \sum (1 - \pi_i)/\pi_i,$$

whereas

$$V(p) = \kappa\pi(1- \pi)$$

and the reduction is as stated. To prove that it is positive it is enough to show that

$$(1 - \pi)/\pi \geq \sum (1 - \pi_i)/\pi_i.$$

This is true for $n = 2$, since, with $\pi = \pi_1\pi_2$,

$$\frac{1 - \pi_1\pi_2}{\pi_1\pi_2} - \frac{1 - \pi_1}{\pi_1} - \frac{1 - \pi_2}{\pi_2} = \frac{(1 - \pi_1)(1 - \pi_2)}{\pi_1\pi_2}$$

and the general case follows by induction similarly.

The improvement using the product can be substantial. For example, if each $\pi_i = \frac{1}{2}$, then $\pi = 2^{-n}$ and the product yields a variance of $\kappa 2^{-2n}n$ against $\kappa 2^{-n}(1 - 2^{-n})$, about $\kappa 2^{-n}$, by direct evaluation. Thus with $n = 4$, the product has one quarter the variance of the direct evaluation.

Many statistical calculations involve both the extension of the conversation and the product law. As we have seen (Section 2.2), the calculation of $p(x_1, x_2, \cdots, x_n)$ is often accomplished by including a parameter that makes the x's conditionally iid, so that

$$p(x_1, x_2, \cdots, x_n) = \int \prod_1^n q(x_i \mid \theta)\pi(\theta)\, d\theta.$$

Since both devices typically involve reductions in variance, the overall improvement can be expected to be substantial. This is borne out by detailed calculations which are not pursued here; they are implicit in the Bayesian calculations in Section 6.7.

6.6 Ratios of Probabilities

Since the product device works so well, it is reasonable to expect that ratios will do badly. A familiar use of a ratio is in the evaluation of a conditional probability $\pi(A \mid B)$ as $\pi(AB)/\pi(B)$. Since $\pi(B)$ is necessarily larger than $\pi(AB)$, there may well be correlation between their two measurements. We therefore consider the indirect assessment of $\pi(A \mid B)$ by $p(AB)/(p(AB) + p(\bar{A}B))$, where AB and $\bar{A}B$ are two terms in a partition with covariance given by Corollary 6.3. Write $a = p(AB)$ and $b = p(\bar{A}B)$ and let $V(a) = \kappa\alpha(1 - \alpha)$, $V(b) = \kappa\beta(1 - \beta)$ and $C(a, b) = -\kappa\alpha\beta$. Using

the delta method,

$$V\left(\frac{a}{a+b}\right) = \frac{\beta^2 V(a) + \alpha^2 V(b) - 2\alpha\beta C(a,b)}{(\alpha+\beta)^4}$$
$$= \frac{\kappa[\beta^2\alpha(1-\alpha) + \alpha^2\beta(1-\beta) + 2\alpha^2\beta^2]}{(\alpha+\beta)^4}$$
$$= \frac{\kappa\alpha\beta}{(\alpha+\beta)^3}.$$

Assuming the direct evaluation of $\pi(A \mid B)$ is governed by the same κ, its variance will be $\kappa\alpha\beta/(\alpha+\beta)^2$. This is smaller than the variance of the indirect method since $\alpha + \beta = \pi(B) < 1$. It therefore pays to assess a conditional probability directly, rather than as a ratio of (unconditional) probabilities.

6.7 Bayes Rule

A famous method of calculating a probability is to use Bayes rule. Its efficacy is now investigated along similar lines to those used for the extension of the conversation, products and ratios; in particular, additive coherence will be assumed. It turns out that the situation is far from clear-cut. The following little argument gives a foretaste of difficulties to come. For me, it first arose in a forensic science application where G is the event that the defendant is truly guilty and E is some evidence before the court. The background information K, including evidence earlier before the court, is omitted from the notation. Bayes rule in log-odds form reads

$$\log\frac{\pi(G \mid E)}{\pi(\bar{G} \mid E)} = \log\frac{\pi(E \mid G)}{\pi(E \mid \bar{G})} + \log\frac{\pi(G)}{\pi(\bar{G})}. \tag{16}$$

Suppose that the likelihoods, $\pi(E \mid G)$ and $\pi(E \mid \bar{G})$, are measured without error. This is often assumed in statistical arguments and is reasonably true for some forensic evidence; for example, with evidence of blood types, $\pi(E \mid G) = 1$ and $\pi(E \mid \bar{G})$ is the frequency of the blood type in the population. In these circumstances, it is immediate from (16) that any error in the odds prior to E will be perpetuated in those posterior to E. On the log-odds scale, Bayes rule does nothing to reduce the variance. If the likelihoods do include error, independent of that in the log-odds, the variance of the latter is actually increased by the additional evidence.

The calculations in the previous sections have not used log-odds but have supposed $V(p \mid \pi) = \kappa\pi(1-\pi)$, implying $V(\log(p/(1-p)) \mid \pi) = \kappa(\pi(1-\pi))^{-1}$. In court cases, hopefully π tends to 0 or 1 so that our method has the variance of log-odds increasing indefinitely. The constancy of variance of log-odds implied by (16)—when the likelihoods are precise—may therefore represent an improvement. The general situation is now investigated more carefully. As far as possible, the notation of Section 6.4 is used because Bayes rule uses the extension in the determination of the normalizing constant.

Consider Bayes rule in the form

$$\pi(B_1 \mid A) = \frac{\pi(A \mid B_1)\pi(B_1)}{\sum \pi(A \mid B_i)\pi(B_i)},$$

where $(B_i: 1 \le i \le n)$ is a partition. As in Section 6.4, write $\alpha_i = \pi(A \mid B_i)$, likelihoods; $\beta_i = \pi(B_i)$, prior; and use the new notation $\gamma_i = \pi(B_i \mid A)$, posterior. The variance of the direct measurement c_1 is to be compared with that of the indirectly obtained value by Bayes rule, $a_1b_1/\sum a_ib_i$. Assumptions 1, 2 and 3 in Section 6.4 still obtain. In addition we use

ASSUMPTION 4. $V(c_i) = \mu\gamma_i(1-\gamma_i)$.

THEOREM. *With Assumptions* 1–4, *the variance of the indirect assessment* $a_1b_1/\sum a_ib_i$, *using Bayes rule, is*

$$\alpha_1\beta_1\{\lambda\alpha_1[(\textstyle\sum'\alpha_i\beta_i)^2 + \beta_1(\sum'\alpha_i^2\beta_i)] + \kappa\beta_1[(1-\alpha_1)(\textstyle\sum'\alpha_i\beta_i)^2 + \alpha_1\textstyle\sum'\beta_i^2\alpha_i(1-\alpha_i)]\}/(\textstyle\sum\alpha_i\beta_i)^4 \tag{17}$$

where $\sum'$ *denotes a summation from* 2 *to* n (*omitting* 1).

The proof is simply a tedious manipulation by the delta method. Write $t_i = a_ib_i$, $t = \sum t_i$, so that the indirect value is t_1/t. Then

$$V(t_1/t) = (\tau^2V(t_1) - 2\tau_1C(t, t_1) + \tau_1^2V(t))\tau\tau^{-4}. \tag{18}$$

Using the delta method again,

$$V(t_1) = \lambda\alpha_1^2\beta_1(1-\beta_1) + \kappa\beta_1^2\alpha_1(1-\alpha_1),$$

$$V(t) = \lambda\textstyle\sum\alpha_i^2\beta_i - \lambda(\sum\alpha_i\beta_i)^2 + \kappa\sum\beta_i^2\alpha_i(1-\alpha_i)$$

and

$$C(t, t_1) = \lambda\alpha_1^2\beta_1(1-\beta_1) - \lambda\alpha_1\beta_1\textstyle\sum'\alpha_i\beta_i + \kappa\beta_1^2\alpha_1(1-\alpha_1).$$

Inserting these three expressions into (18) and making some rearrangements of terms, we have the result stated, (17).

This variance has to be compared with that obtained directly, namely

$$V(c_1) = \mu\gamma_1(1-\gamma_1) = \frac{\mu\alpha_1\beta_1(\sum'\alpha_i\beta_i)}{(\sum\alpha_i\beta_i)^2}. \tag{19}$$

There is no simple relationship between (17) and (19), and numerical work demonstrates that it is possible for either to be the smaller even when $\kappa = \lambda = \mu$. To obtain a feel for what is happening, we take two special cases: first, a partition into two events, $n = 2$, as

in the court case, $B_1 = G$, $B_2 = \bar{G}$: second, a fine partition with n large and the largest β_i tending to 0 (see Section 6.4).

With $n = 2$, (17) reduces to

$$V(t_1/t) = \frac{\alpha_1\beta_1\alpha_2\beta_2(\lambda\alpha_1\alpha_2 + \kappa\beta_1\beta_2(\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2))}{(\alpha_1\beta_1 + \alpha_2\beta_2)^4}$$

and

$$V(c_1) = \frac{\mu\alpha_1\beta_1\alpha_2\beta_2}{(\alpha_1\beta_1 + \alpha_2\beta_2)^2}.$$

Detailed comparison must rest on the individual values of κ, λ and μ. Bayes rule will be favoured if the likelihoods are precisely determined, $\kappa = 0$, as in some forensic cases. If, in addition, the priors and posteriors are equally precisely assessed, $\lambda = \mu$, Bayes will be superior iff

$$\alpha_1\alpha_2 < (\alpha_1\beta_1 + \alpha_2\beta_2)^2$$

or

$$(\alpha_1\alpha_2)^{1/2} < \alpha_1\beta_1 + \alpha_2(1 - \beta_1)$$

on remembering $\beta_2 = 1 - \beta_1$. Then Bayes is better iff

$$\beta_1 > (<)\ \alpha_2^{1/2}/(\alpha_1^{1/2} + \alpha_2^{1/2})$$

when $\alpha_1 > (<)\ \alpha_2$.

In words, in considering the event B_1 of higher (lower) likelihood, Bayes will be better if it has sufficiently high (low) prior probability. Essentially, for Bayes to be effective there has to be an agreement between likelihood and prior.

The case of large n is easier to appreciate. If each β_i tends to 0, the dominant term in (17) is the first and it reduces to $\lambda\alpha_1^2\beta_1/(\sum \alpha_i\beta_i)^2 + o(\beta_1)$. The direct assessment (19) similarly reduces to a variance of $\mu\alpha_1\beta_1/\sum \alpha_i\beta_i + o(\beta_1)$. Consequently Bayes is preferred iff

$$\alpha_1 < (\mu/\lambda)(\sum \alpha_i\beta_i).$$

In the language and notation of Section 6.4, $\sum \alpha_i\beta_i = E(\alpha)$. Consequently Bayes is better only if the likelihood is less than a multiple, μ/λ, of the 'average' likelihood, the average being with respect to the prior. Notice that the precision for the likelihoods, described through κ, is irrelevant for large n.

6.8 Predictive Bayes

Bayes rule in its parametric form (thinking of the B's as corresponding to the values of a parameter) does not necessarily appear as a good measuring device at least with the rather restrictive assumptions made in the above analysis. However, it was argued in Section 2.1 that the basic problem of inference is really prediction from past values x to future ones y and that parameters are merely a device introduced to simplify the calculations (Section 2.2). An analysis similar to that performed in Section 6.7 is now carried out for Bayes rule in its predictive form. The notation is slightly changed. The partition remains (B_i) but the event A corresponding to observed data is replaced by X and Y corresponds to future data. The formula is then

$$\pi(Y \mid X) = \frac{\sum \pi(Y \mid B_i)\pi(X \mid B_i)\pi(B_i)}{\sum \pi(X \mid B_i)\pi(B_i)}$$

on assuming X and Y independent given the partition. As before write $\alpha_i = \pi(X \mid B_i)$, $\beta_i = \pi(B_i)$, but introduce the notation $\lambda_i = \pi(Y \mid B_i)$. The first task is to determine the variance of $\sum l_i a_i b_i / \sum a_i b_i$. Assumptions 1–3 of Section 6.4 still obtain. In addition:

ASSUMPTION 4. $V(l_i) = \nu\lambda_i(1 - \lambda_i)$, $C(l_i, a_j) = C(l_i, b_j) = 0$.

The calculations are even more tedious and complicated than in the parametric situation. We shall therefore confine the analysis to the case of large n with the largest probability in the partition going to 0 (Section 6.4). In that case it was found convenient to write expressions like $\sum \alpha_i\beta_i$ as $E(\alpha)$ for a random quantity α taking values α_i with probabilities β_i. Using this convention, we have

THEOREM. *With Assumptions 1–4, the variance of the predictive assessment* $\sum l_i a_i b_i / \sum a_i b_i$ *is approximately*

$$\lambda(E(\alpha^2\Lambda^2) - E(\alpha\Lambda)^2)/E(\alpha)^4. \tag{20}$$

Here Λ *is a random quantity taking values* $E(\alpha)\lambda_i - E(\lambda\alpha) = \Lambda_i$ *with probabilities* β_i.

Apply the delta method to $\sum l_i a_i b_i / \sum a_i b_i$ and replace any terms that do not involve δ's by their expectations (thus $\sum a_i b_i$ is replaced by $E(\alpha)$). We have

$$\begin{aligned}
&\delta(\textstyle\sum l_i a_i b_i / \sum a_i b_i) \\
&\quad = [(\textstyle\sum a_i b_i)(\sum l_i a_i \delta b_i + \sum l_i b_i \delta a_i + \sum a_i b_i \delta l_i) \\
&\qquad\qquad - (\textstyle\sum l_i a_i b_i)(\sum a_i \delta b_i + \sum b_i \delta a_i)]/(\sum a_i b_i)^2 \\
&\quad = (E(\alpha) \textstyle\sum a_i b_i \delta l_i + \sum b_i L_i \delta a_i + \sum a_i L_i \delta b_i)/E(\alpha)^2
\end{aligned}$$

with $L_i = E(\alpha)l_i - E(\lambda\alpha)$. Squaring and taking expectations

$$\begin{aligned}
&V(\textstyle\sum l_i a_i b_i / \sum a_i b_i) \\
&= \Big(E(\alpha)^2 \textstyle\sum \alpha_i^2\beta_i^2\nu\lambda_i(1 - \lambda_i) + \sum \beta_i^2\Lambda_i^2\kappa\alpha_i(1 - \alpha_i) \\
&\qquad + \textstyle\sum \alpha_i^2\Lambda_i^2\lambda\beta_i(1 - \beta_i) - \sum_{i \neq j} \alpha_i\alpha_j L_i L_j \beta_i\beta_j\Big) \Big/ E(\alpha)^4.
\end{aligned}$$

The first two terms are of small order and may be neglected. The last two can be written in the form given in the statement of the theorem. Notice that the result does not depend on the constants κ and ν governing the errors in the two likelihoods $\pi(X \mid B_i)$ and $\pi(Y \mid B_i)$ but only on λ, corresponding to measurement of the prior.

This variance has to be compared with that obtained by a direct measurement of $\pi(Y \mid X) = E(\alpha\lambda)/E(\alpha) = \gamma$ say.

ASSUMPTION 5.

$$V(p(Y \mid X)) = \lambda\gamma(1 - \gamma). \tag{21}$$

This means that the posterior predictive and the prior, both measured directly, have the same measurement errors.

Again, neither direct nor indirect method always has the advantage. Extensive numerical work shows that the variance (20) by use of Bayes is usually substantially smaller than that obtained directly (21). It proved quite difficult to find a contrary case but it can be found when $E(\alpha\lambda) = E(\alpha)E(\lambda)$: that is, α and λ are uncorrelated with respect to (B_i). This would be an unusual case since X and Y ordinarily refer to similar things and would be expected to be correlated in this sense. Thus Bayes formula, in its predictive role, does appear to be ordinarily efficacious. This is not surprising since it uses the extension of the conversation, which has been seen to be effective (Section 6.4), in both numerator and denominator, unlike parametric Bayes which only uses it in the latter. Presumably predictive Bayes is only inadequate because of its use of a ratio (Section 6.6).

6.9 Summary

To be useful, probability theory has to be supplemented by a method of measurement. The measurement can either be direct or indirect, using formulae from the theory. Several such formulae have been studied using a very special error structure that incorporates additive coherence. The results show that extension of the conversation and the product rule are both useful measurement devices in that they can reduce the error. Evaluation of a conditional probability as the ratio of unconditionals is unsatisfactory. Bayes formula in its parametric form can sometimes be useful but often does not result in an improvement. In its predictive form though, it ordinarily does give a better measurement, at least for fine partitions.

My personal conviction is that some calculus of probability measurement, analogous to least squares, is essential. Whether the methods described here, particularly additive coherence and the conclusions, especially about Bayes rule, provide even the crude basis of such a calculus, is more doubtful.

ACKNOWLEDGMENTS

This work was sponsored by the Office of Naval Research, Contract N00014-85-C-0268 to Decision Science Consortium, Inc., of Reston, Virginia. Some of the writing was carried out at the Division of Statistics on the Davis campus whilst in receipt of a Regents' Professorship of the University of California. I am grateful to Rex V. Brown and George Roussas respectively for these arrangements.

REFERENCES

ALLAIS, M. (1987). Allais's paradox. *The New Palgrave: A Dictionary of Economics* **1** 80–82. MacMillan, London.

ANDERSON, T. W. (1987). Comment on "A review of multivariate analysis" by M. J. Schervish. *Statist. Sci.* **2** 413–417.

ANSCOMBE, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58** 365–383.

BARTHOLOMEW, D. J. (1967). Hypothesis testing when the sample size is treated as a random variable (with discussion). *J. Roy. Statist. Soc. Ser. B* **29** 53–82.

BASU, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A* **26** 3–16.

BASU, D. (1988). *Statistical Information and Likelihood.* Springer, New York.

BATHER, J. A. (1985). On the allocation of treatments in sequential medical trials (with discussion). *Internat. Statist. Rev.* **53** 1–13.

BAYARRI, M. J., DEGROOT, M. H. and KADANE, J. B. (1988). What is the likelihood function? (with discussion). In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 3–27. Springer, New York.

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317–352.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle.* IMS, Hayward, Calif.

BERRY, D. A. (1988). Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 79–94. Clarendon Press, Oxford.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

COHEN, L. (1958). On mixed single sample experiments. *Ann. Math. Statist.* **29** 947–971.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics* **25** 617–657.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

DE FINETTI, B. (1937). Le prévision ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7** 1–68. (Translated in *Studies in Subjective Probability* (H. E. Kyburg, Jr. and H. E. Smokler, eds.) 95–158. Wiley, New York, 1964.)

DE FINETTI, B. (1974/5). *Theory of Probability.* **1**, **2**. Wiley, New York.

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

DEGROOT, M. H. and FIENBERG, S. E. (1986). Comparing probability forecasters: Basic binary concepts and multivariate extensions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) 247–264. North-Holland, Amsterdam.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

DICKEY, J. M., DAWID, A. P. and KADANE, J. B. (1986). Subjective probability assessment methods for multivariate-t and matrix-t models. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) 177–195. North-Holland, Amsterdam.

DURBIN, J. (1987). Statistics and statistical science. *J. Roy. Statist. Soc. Ser. A* **150** 177–191.

EVETT, I. W. (1984). A quantitative theory for interpreting transfer evidence in criminal cases. *Appl. Statist.* **33** 25–32.

FINKELSTEIN, M. O. (1978). *Quantitative Methods in Law.* Free Press, New York.

FISHBURN, P. C. (1986). The axioms of subjective probability (with discussion). *Statist. Sci.* **1** 335–358.

GEISSER, S. (1985). On the prediction of observables: A selective update (with discussion). In *Bayesian Statistics 2.* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 203–299. North-Holland, Amsterdam.

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31** 423–447.

HUZURBAZAR, V. S. (1955). On the certainty of an inductive inference. *Proc. Cambridge Philos. Soc.* **51** 761–762.

JEFFREYS, H. (1939). *Theory of Probability.* Clarendon Press, Oxford.

JEWELL, W. S. (1974). Credible means are exact Bayesian for simple exponential families. *Astin Bull.* **8** 77–90.

KADANE, J. B., DICKEY, J. M., WINKLER, R. L., SMITH, W. S. and PETERS, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75** 845–854.

KADANE, J. B. and WINKLER, R. L. (1988). Separating probability elicitation from utilities. *J. Amer. Statist. Assoc.* **83** 357–363.

KAHNEMAN, D., SLOVIC, P. and TVERSKY, A. (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge Univ. Press, Cambridge.

KOLMOGOROV, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin. English edition, Chelsea, New York, 1950.

KUHN, T. S. (1974). *Structure of Scientific Revolutions.* Univ. Chicago Press, Chicago.

LAVIS, D. A. and MILLIGAN, P. J. (1985). The work of E. T. Jaynes on probability, statistics and statistical physics. *Brit. J. Philos. Sci.* **36** 193–210.

LINDLEY, D. V. (1972). *Bayesian Statistics: A Review.* SIAM, Philadelphia.

LINDLEY, D. V. (1984). Prospects for the future: The next 50 years (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 359–367.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, Mass.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MEINHOLD, R. J. and SINGPURWALLA, N. D. (1983). Understanding the Kalman filter. *Amer. Statist.* **37** 123–127.

MOSTELLER, F. (1988). Broadening the scope of statistics and statistical education. *Amer. Statist.* **42** 93–99.

O'HAGAN, A. (1988). *Probability: Methods and Measurement.* Chapman and Hall, London.

PRATT, J. W., RAIFFA, H. and SCHLAIFER, R. (1964). The foundations of decision under uncertainty: An elementary exposition. *J. Amer. Statist. Assoc.* **59** 353–375.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Harvard Univ. Press, Cambridge, Mass.

RAMSEY, F. P. (1931). Truth and probability. In *The Foundations of Mathematics and Other Logical Essays.* Kegan, Paul, Trench, Trubner, London.

ROBINSON, G. K. (1987). That BLUP is a good thing—the estimation of random effects. CSIRO, Melbourne.

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

SHAFER, G. (1976). *A Mathematical Theory of Evidence.* Princeton Univ. Press, Princeton, N.J.

SHAFER, G. (1986). Savage revisited (with discussion). *Statist. Sci.* **1** 463–501.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. Ser. B* **23** 1–37.

STURROCK, P. A. (1973). Evaluation of astrophysical hypotheses. *Astrophys. J.* **182** 569–580.

WALD, A. (1947). *Sequential Analysis.* Wiley, New York.

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

WALD, A. and WOLFOWITZ, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19** 326–339.

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Assoc.* **80** 73–97.

WINKLER, R. L. (1986). On "good probability appraisers." In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) 265–278. North-Holland, Amsterdam.

ZADEH, L. A. (1983). The role of fuzzy logic in the measurement of uncertainty in expert systems. *Fuzzy Sets and Systems* **11** 199–227.

Comment

George A. Barnard

The invitation to comment on Lindley's lectures arrived with a close deadline at a busy time. But, as always, his style is so clear and his thought so bold that I find the temptation to discuss at least some of his points irresistible.

When the University of Oxford—"the home of lost causes"—at last decided to set up a lectureship in mathematical statistics they called the resulting group LIDASE: the lectureship in the design and analysis of

George A. Barnard is Professor Emeritus at the University of Essex. His mailing address is Mill House, 54 Hurst Green, Brightlingsea, Colchester, Essex CO7 0EH, England.