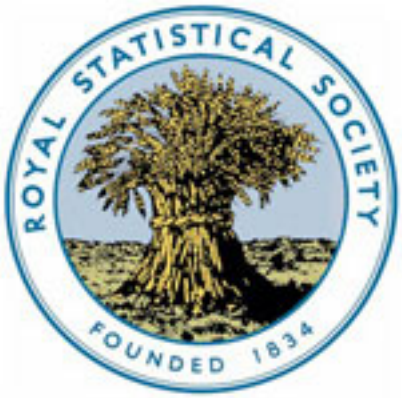

The Statistician (2000)
49, *Part* 3, *pp.* 293–337

The philosophy of statistics

Dennis V. Lindley

Minehead, UK

[Received June 1999]

Summary. This paper puts forward an overall view of statistics. It is argued that statistics is the study of uncertainty. The many demonstrations that uncertainties can only combine according to the rules of the probability calculus are summarized. The conclusion is that statistical inference is firmly based on probability alone. Progress is therefore dependent on the construction of a probability model; methods for doing this are considered. It is argued that the probabilities are personal. The roles of likelihood and exchangeability are explained. Inference is only of value if it can be used, so the extension to decision analysis, incorporating utility, is related to risk and to the use of statistics in science and law. The paper has been written in the hope that it will be intelligible to all who are interested in statistics.

Keywords: Conglomerability; Data analysis; Decision analysis; Exchangeability; Law; Likelihood; Models; Personal probability; Risk; Scientific method; Utility

1. Introduction

Instead of discussing a specific problem, this paper provides an overview within which most statistical issues can be considered. 'Philosophy' in the title is used in the sense of 'The study of the general principles of some particular branch of knowledge, experience or activity' (Onions, 1956). The word has recently acquired a reputation for being concerned solely with abstract issues, divorced from reality. My intention here is to avoid excessive abstraction and to deal with practical matters concerned with our subject. If the practitioner who reads this paper does not feel that the study has benefited them, my writing will have failed in one of its endeavours. The paper tries to develop a way of looking at statistics that will help us, as statisticians, to develop better a sound approach to any problem which we might encounter. Technical matters have largely been avoided, not because they are not important, but in the interests of focusing on a clear understanding of how a statistical situation can be studied. At some places, matters of detail have been omitted to highlight the key idea. For example, probability densities have been used without an explicit mention of the dominating measure to which they refer.

The paper begins by recognizing that statistical issues concern uncertainty, going on to argue that uncertainty can only be measured by probability. This conclusion enables a systematic account of inference, based on the probability calculus, to be developed, which is shown to be different from some conventional accounts. The likelihood principle follows from the basic role played by probability. The role of data analysis in the construction of probability models and the nature of models are next discussed. The development leads to a method of making decisions and the nature of risk is considered. Scientific method and its application to some legal issues are explained within the probabilistic framework. The conclusion is that we have here a satisfactory general set of statistical procedures whose implementation should improve statistical practice.

Address for correspondence: Dennis V. Lindley, "Woodstock", Quay Lane, Minehead, Somerset, TA24 5QU, UK.
E-mail: thombayes@aol.com

 0039–0526/00/49293

The philosophy here presented places more emphasis on model construction than on formal inference. In this it agrees with much recent opinion. A reason for this change of emphasis is that formal inference is a systematic procedure within the calculus of probability. Model construction, by contrast, cannot be so systematic.

The paper arose out of my experiences at the Sixth Valencia Conference on Bayesian Statistics, held in June 1998 (Bernardo *et al.*, 1999). Although I was impressed by the overall quality of the papers and the substantial advances made, many participants did not seem to me fully to appreciate the Bayesian philosophy. This paper is an attempt to describe my version of that philosophy. It is a reflection of 50 years' statistical experience and a personal change from a frequentist, through objective Bayes, to the subjective attitude presented here. No attempt has been made to analyse in detail alternative philosophies, only to indicate where their conclusions differ from those developed here and to contrast the resulting practical methods.

2. Statistics

To discuss the philosophy of statistics, it is necessary to be reasonably clear what it is the philosophy of, not in the sense of a precise definition, so that this is 'in', that is 'out', but merely to be able to perceive its outlines. The suggestion here is that statistics is the study of uncertainty (Savage, 1977): that statisticians are experts in handling uncertainty. They have developed tools, like standard errors and significance levels, that measure the uncertainties that we might reasonably feel. A check of how well this description of our subject agrees with what we actually do can be performed by looking at the four series of journals published by the Royal Statistical Society in 1997. These embrace issues as diverse as social accounting and stable laws. Apart from a few exceptions, like the algorithms section (which has subsequently been abandoned) and a paper on education, all the papers deal either directly with uncertainty or with features, like stable laws, which arise in problems that exhibit uncertainty. Support for this view of our subject is provided by the fact that statistics plays a greater role in topics that have variability, giving rise to uncertainty, as an essential ingredient, than in more precise subjects. Agriculture, for example, enjoys a close association with statistics, whereas physics does not. Notice that it is only the manipulation of uncertainty that interests us. We are not concerned with the matter that is uncertain. Thus we do not study the mechanism of rain; only whether it will rain. This places statistics in a curious situation in that we are, as practitioners, dependent on others. The forecast of rain will be dependent on both meteorologist and statistician. Only as theoreticians can we exist alone. Even there we suffer if we remain too divorced from the science. The term 'client' will be used in reference to the person, e.g. scientist or lawyer, who encounters uncertainty in their field of study.

The philosophical position adopted here is that statistics is essentially the study of uncertainty and that the statistician's role is to assist workers in other fields, the clients, who encounter uncertainty in their work. In practice, there is a restriction in that statistics is ordinarily associated with data; and it is the link between the uncertainty, or variability, in the data and that in the topic itself that has occupied statisticians. Some writers even restrict the data to be frequency data, capable of near-identical repetition. Uncertainty, away from data, has rarely been of statistical interest. Statisticians do not have a monopoly of studies of uncertainty. Probabilists discuss how randomness in one part of a system affects other parts. Thus the model for a stochastic process provides predictions about the data that the process will provide. The passage from process to data is clear; it is when we attempt a reversal and go from data to process that difficulties appear. This paper is mainly devoted to this last phase, commonly called inference, and the action that it might generate.

Notice that uncertainty is everywhere, not just in science or even in data. It provides a motivation for some aspects of theology (Bartholomew, 1988). Therefore, the recognition of statistics as uncertainty would imply an extensive role for statisticians. If a philosophical position can be developed that embraces all uncertainty, it will provide an important advance in our understanding of the world. At the moment it would be presumptive to claim so much.

3. Uncertainty

Acceptance that statistics is the study of uncertainty implies that it is necessary to investigate the phenomenon. A scientific approach would mean the measurement of uncertainty; for, to follow Kelvin, it is only by associating numbers with any scientific concept that the concept can be properly understood. The reason for measurement is not just to make more precise the notion that we are more uncertain about the stock-market than about the sun rising tomorrow, but to be able to combine uncertainties. Only exceptionally is there one element of uncertainty in a problem; more realistically there are several. In the collection of data, there is uncertainty in the sampling unit, and then in the number reported in the sampling. In an archetypal statistical problem, there is uncertainty in both data and parameter. The central problem is therefore the combination of uncertainties. Now the rules for the combination of numbers are especially simple. Furthermore, numbers combine in two ways, addition and multiplication, so leading to a richness of ideas. We want to measure uncertainties in order to combine them. A politician said that he preferred adverbs to numbers. Unfortunately it is difficult to combine adverbs.

How is this measurement to be achieved? All measurement is based on a comparison with a standard. For length we refer to the orange–red line of the krypton-86 isotope. The key role of comparisons means that there are no absolutes in the world of measurement. This is a point which we shall return to in Section 11. It is therefore necessary to find a standard for uncertainty. Several have been suggested but the simplest is historically the first, namely games of chance. These provided the first uncertainties to be studied systematically. Let us therefore use as our standard a simple game.

Consider before you an urn containing a known number N of balls that are as nearly identical as modern engineering can make them. Suppose that one ball is drawn at random from the urn. For this to make sense, it is needful to define randomness. Imagine that the balls are numbered consecutively from 1 to N and suppose that, at no cost to you, you were offered a prize if ball 57 were drawn. Suppose, alternatively, that you were offered the same prize if ball 12 were drawn. If you are indifferent between the two propositions and, in extension, beween any two numbers between 1 and N, then, for you, the ball is drawn at random. Notice that the definition of randomness is subjective; it depends on you. What is random for one person may not be random for another. We shall return to this aspect in Section 8.

Having said what is meant by the drawing of a ball at random, forget the numbers and suppose that R of the balls are red and the remainder white, the colouring not affecting your opinion of randomness. Consider the uncertain event that the ball, withdrawn at random, is red. The suggestion is that this provides a standard for uncertainty and that the measure is R/N, the proportion of red balls in the urn. There is nothing profound here, being just a variant of the assumption on which games of chance are based. Now pass to any event, or proposition, which can either happen or not, be true or false. It is proposed to measure your uncertainty associated with the event happening by comparison with the standard. If you think that the event is just as uncertain as the random drawing of a red ball from an urn containing N balls, of which R are red, then the event has uncertainty R/N for you. R and N are for you to choose. For given N, it is easy to see that there cannot be more than one such R. There is now a measure of uncertainty for any

event or proposition. Before proceeding, let us consider the measurement process carefully.

A serious assumption has been made that the concept of uncertainty can be isolated from other features. In discussing randomness, it was useful to compare a prize given under different circumstances, ball 57 or ball 12. Rewards will not necessarily work in the comparison of an event with a standard. For example, suppose that the event whose uncertainty is being assessed is the explosion of a nuclear weapon, within 50 miles of you, next year. Then a prize of £10000, say, will be valued differently when a red ball appears from when it will when fleeing from the radiation. The measurement process just described assumes that you can isolate the uncertainty of the nuclear explosion from its unpleasant consequences. For the moment we shall make the assumption, returning to it in Section 17 to show that, in a sense, it is not important. Ramsey (1926), whose work will be discussed in Section 4, introduced the concept of an 'ethically neutral event' for which the comparison with the urn presents fewer difficulties. Nuclear bombs are not ethically neutral.

In contrast, notice an assumption that has *not* been made. For any event, including the nuclear bomb, it has not been assumed that you can do the measurement to determine R (and N, but that only reflects the precision in your assessment of the uncertainty). Rather, we assume that you would wish to do it, were you to know how. All that is assumed of any measurement process is that it is reasonable, not that it can easily be done. Because you do not know how to measure the distance to our moon, it does not follow that you do not believe in the existence of a distance to it. Scientists have spent much effort on the accurate determination of length because they were convinced that the concept of distance made sense in terms of krypton light. Similarly, it seems reasonable to attempt the measurement of uncertainty.

4. Uncertainty and probability

It has been noted that a prime reason for the measurement of uncertainties is to be able to combine them, so let us see how the method suggested accomplishes this end. Suppose that, of the N balls in the urn, R are red, B are blue and the remainder white. Then the uncertainty that is associated with the withdrawal of a coloured ball is $(R + B)/N = R/N + B/N$, the sum of the uncertainties associated with red, and with blue, balls. The same result will obtain for any two exclusive events whose uncertainties are respectively R/N and B/N and we have an addition rule for your uncertainties of exclusive events.

Next, suppose again that R balls are red and the remaining $N - R$ white; at the same time, S are spotted and the remaining $N - S$ plain. The urn then contains four types of ball, of which one type is both spotted and red, of which the number is say T. Then the uncertainty associated with the withdrawal of a spotted red ball is T/N, which is equal to $R/N \times T/R$, the product of the uncertainty of a red ball and that of spotted balls among the red. Again the same result will apply for any two events being compared with coloured and with spotted balls and we have a product rule for uncertainties.

The addition and product rules just obtained, together with the convexity rule that the measurement R/N always lies in the (convex) unit interval, are the defining rules of probability, at least for a finite number of events (see Section 5). The conclusion is therefore that the measurements of uncertainty can be described by the calculus of probability. In the reverse direction, the rules of probability reduce simply to the rules governing proportions. Incidentally, this helps to explain why frequentist arguments are often so useful: the combination of uncertainties can be studied by proportions, or frequencies, in a group, here of balls. The mathematical basis of probability is very simple and it is perhaps surprising that it yields complicated and useful results.

The conclusions are that statisticians are concerned with uncertainty and that uncertainty can

be measured by probability. The sketchy demonstration here may not be thought adequate for such an important conclusion, so let us look at other approaches.

Historically, uncertainty has been associated with games of chance and gambling. Hence one way of measuring uncertainty is through the gambles that depend on it. The willingness to stake s on an event to win w if the event occurs is, in effect, a measure of the uncertainty of the event expressed through the odds (against) of w/s to 1. The combination of events is now replaced by collections of gambles. To fix ideas, contemplate the situation in horse-racing. With two horses running in the *same* race, bets on them separately may be considered as well as a bet on *either* of them winning. With horses in *different* races, the event of *both* winning may be used. Using combinations like these, the concept of what in Britain we term a Dutch book can be employed. A series of bets at specified odds is said to constitute a Dutch book if it is possible to place a set of stakes in such a way that one is sure to come out overall with an increase in assets. A bookmaker never states odds for which a Dutch book may be made. It is easy to show (de Finetti, 1974, 1975) that avoidance of a Dutch book is equivalent to the probabilities, derived from the odds, obeying the convexity, addition and multiplication rules of probability already mentioned. For odds o, the corresponding probability is $p = (1 + o)^{-1}$. In summary, the use of odds, combined with the impossibility of a Dutch book, leads back to probability as before.

de Finetti introduced another approach. Suppose that you are required to state your numerical uncertainty for an event, and you are told that you will be scored by an amount $S(x, E)$ if you state x, where $E = 1$ or $E = 0$ if the event is subsequently shown to be true or false respectively. For different events, the scores are to be added. A possible penalty score used by him is $S(x, E) = (x - E)^2$, but, aside from a few scores that lead to ridiculous conclusions, any function will do. Then if you use x as some function of probability, i.e. as a function of something that obeys those three rules of probability, you will be sure, in the sense of obtaining a smaller penalty score, to do better than someone who acts in another way. Which function depends on S. This approach is attractive because it is operational and has been used to train weather forecasters. Generally it provides an empirical check on the quality of your probability assessments and hence a test of your abilities as a statistician (see Section 9). The important point here is that again it leads to probability. We next study a third group of methods that again compels the use of probability to describe uncertainty.

The Royal Statistical Society, together with many other statistical groups, was originally set up to gather and publish data. This agrees with our argument about uncertainty because part of the purpose behind the data gathering was a reduction in uncertainty. It remains an essential part of statistical activity today and most Governments have statistical offices whose function is the acquisition and presentation of statistics. It did not take long before statisticians wondered how the data might best be used and modern statistical inference was born. Then statisticians watched as the inferences were used as the basis of decisions, or actions, and began, with others, to concern themselves with decision-making. It is now explained how this leads again to probability.

The first person to make the step to decision-making seems to have been Ramsey (1926). He asked the simple question 'how should we make decisions in the face of uncertainty?'. He made some apparently reasonable assumptions and from them deduced theorems. The theorem that concerns us here is that which says that the uncertainty measures should obey the rules of the probability calculus. So we are back to probability again. Ramsey's work was unappreciated until Savage (1954) asked the same question and from different, but related, assumptions came up with the same result, namely that it has to be probability. Savage also established a link with de Finetti's ideas. Since then, many others have explored the field with similar results. My personal favourite among presentations that have full rigour is that of DeGroot (1970), chapter 6. An excellent recent presentation is Bernardo and Smith (1994).

5. Probability

The conclusion is that measurements of uncertainty must obey the rules of the probability calculus. Other rules, like those of fuzzy logic or possibility theory, dependent on maxima and minima, rather than sums and products, are out. So are some rules used by statisticians; see Section 6. All these derivations, whether based on balls in urns, gambles, scoring rules or decision-making, are based on assumptions. Since these assumptions imply such important results, it is proper that they are examined with great care. Unfortunately, the great majority of statisticians do not do this. Some deny the central result about probability, without exploring the reasons for it. It is not the purpose of this paper to provide a rigorous account but let us look at one assumption because of its later implications. As with all the assumptions, it is intended to be self-evident and that you would feel foolish if you were to be caught violating it. It is based on a primitive notion of one event being more likely than another. ('Likely' is not used in its technical sense, but as part of normal English usage.) We write 'A is more likely than B' as $A > B$. The assumption is that if A_1 and A_2 are exclusive, and the same is true of B_1 and B_2, then $A_i > B_i$ $(i = 1, 2)$ imply $A_1 \cup A_2 > B_1 \cup B_2$. ($A_1 \cup A_2$ means the event that is true whenever either A_1 or A_2 is true.) We might feel unhappy if we thought that the next person to pass through a door was more likely to be a non-white female, A_1, than a non-white male, B_1, that a white female, A_2, is more likely than a white male, B_2, yet a female, $A_1 \cup A_2$ was *less* likely than a male, $B_1 \cup B_2$. The developments outlined above start from assumptions, or axioms, of this character. The important point is that they all lead to probability being the only satisfactory expression of uncertainty.

The last sentence is not strictly true. Some writers have considered the axioms carefully and produced objections. A fine critique is Walley (1991), who went on to construct a system that uses a pair of numbers, called upper and lower probabilities, in place of the single probability. The result is a more complicated system. My position is that the complication seems unnecessary. I have yet to meet a situation in which the probability approach appears to be inadequate and where the inadequacy can be fixed by employing upper and lower values. The pair is supposed to deal with the precision of probability assertions; yet probability alone contains a measure of its own precision. I believe in simplicity; provided that it works, the simpler is to be preferred over the complicated, essentially Occam's razor.

With the conclusion that uncertainty is only satisfactorily described by probability, it is convenient to state formally the three rules, or axioms, of the probability calculus. Probability depends on two elements: the uncertain event and the conditions under which you are considering it. In the extraction of balls from an urn, your probability for red depends on the condition that the ball is drawn at random. We write $p(A|B)$ for your probability of A when you know, or are assuming, B to be true, and we speak of your probability of A, given B. The rules are as follows.

(a) Rule 1 (convexity): for all A and B, $0 \leqslant p(A|B) \leqslant 1$ and $p(A|A) = 1$.
(b) Rule 2 (addition): if A and B are exclusive, given C,

$$p(A \cup B|C) = p(A|C) + p(B|C).$$

(c) Rule 3 (multiplication): for all A, B and C,

$$p(AB|C) = p(A|BC)\, p(B|C).$$

(Here AB means the event that occurs if, and only if, both A and B occur.) The convexity rule is sometimes strengthened to include $p(A|B) = 1$ only if A is a logical consequence of B. The addition is called Cromwell's rule.

There is one point about the addition rule that appears to be merely a mathematical nicety but in fact has important practical consequences to be exhibited in Section 8. With the three rules as

stated above, it is easy to extend the addition rule for two, to any finite number of events. None of the many approaches already discussed lead to the rule's holding for an infinity of events. It is usual to suppose that it does so hold because of the undesirable results that follow without it. This can be made explicit either by simply restating the addition rule, or by adding a fourth rule which, together with the three above, leads to addition for an infinity of exclusive events. My personal preference is for the latter and to add the following.

(d) Rule 4 (conglomerability): if $\{B_n\}$ is a partition, possibly infinite, of C and $p(A|B_nC) = k$, the same value for all n, then $p(A|C) = k$.

(It is easy to verify that rule 4 follows from rules 1–3 when the partition is finite. The definition is due to de Finetti.) Conglomerability is in the spirit of a class of rules known as 'sure things'. Roughly, if whatever happens (whatever B_n) your belief is k, then your belief is k unconditionally. The assumption described earlier in this section is in the same spirit. Some statisticians appear to be conglomerable only when it suits them: hence the practical connection to be studied in Section 8. Note that the rules of probability are here not stated as axioms in the manner found in texts on probability. They are deductions, apart from rule 4, from other, more basic, assumptions.

6. Significance and confidence

The reaction of many statisticians to the assertion that they should use probability will be to say that they do it already, and that the developments here described do nothing more than give a little cachet to what is already being done. The journals are full of probabilities: normal and binomial distributions abound; the exponential family is everywhere. It might even be claimed that no other measure of uncertainty is used: few, if any, statisticians embrace fuzzy logic. Yet this is not true; statisticians do use measures of uncertainty that do *not* combine according to the rules of the probability calculus.

Consider a hypothesis H, that a medical treatment is ineffectual, or that a specific social factor does not influence crime levels. The physician, or sociologist, is uncertain about H, and data are collected in the hope of removing, or at least reducing, the uncertainty. A statistician called in to advise on the uncertainty aspect may recommend that the client uses, as a measure of uncertainty, a tail area, significance level, with H as the null hypothesis. That is, assuming that H is true, the probability of the observed, or more extreme, data is calculated. This is a measure of the credence that can be put on H; the smaller the probability, the smaller is the credence.

This usage flies in the face of the arguments above which assert that uncertainty about H needs to be measured by a probability for H. A significance level is *not* such a probability. The distinction can be expressed starkly:

significance level—the probability of some aspect of the data, given H is true;
probability—your probability of H, given the data.

The prosecutor's fallacy is well known in legal circles. It consists in confusing $p(A|B)$ with $p(B|A)$, two values which are only rarely the same. The distinction between significance levels and probability is almost the prosecutor's fallacy: 'almost' because although B, in the prosecutor form, may be equated with H, the data are treated differently. Probability uses A as data. Adherents of significance levels soon recognized that they could not use just the data but had to include 'more extreme' data in the form of the tail of a distribution. As Jeffreys (1961) put it: the level includes data which might have happened but did not.

From hypothesis testing, let us pass to point estimation. A parameter θ might be the effect of

the medical treatment, or the influence of the social factor on crime. Again θ is uncertain, data might be collected and a statistician consulted (hopefully not in that order). The statistician will typically recommend a confidence interval. The development above based on measured uncertainty will use a probability density for θ, and perhaps an interval of that density. Again we have a contrast similar to the prosecutor's fallacy:

confidence—probability that the interval includes θ;
probability—probability that θ is included in the interval.

The former is a probability statement about the interval, given θ; the latter about θ, given the data. Practitioners frequently confuse the two. More important than the confusion is the fact that neither significance levels nor statements of confidence combine according to the rules of the probability calculus.

Does the confusion matter? At a theoretical level, it certainly does, because the use of any measure that does not combine according to the rules of the probability calculus will ultimately violate some of the basic assumptions that were intended to be self-evident and to cause embarrassment if violated. At a practical level, it is not so clear and it is necessary to spend a while explaining the practical implications. Statisticians tend to study problems in isolation, with the result that combinations of statements are not needed, and it is in the combinations that difficulties can arise, as was seen in the colour–sex example in Section 5. For example, it is rarely possible to make a Dutch book against statements of significance levels. Some common estimators are known to be inadmissible. The clearest example of an important violation occurs with the relationship between a significance level and the sample size n on which it is based. The interpretation of 'significant at 5%' depends on n, whereas a probability of 5% always means the same. Statisticians have paid inadequate attention to the relationships between statements that they make and the sample sizes on which they are based. There are theoretical reasons (Berger and Delampady, 1987) for thinking that it is too easy to obtain 5% significance. If so, many experiments raise false hopes of a beneficial effect that does not truly exist.

Individual statistical statements, made in isolation, may not be objectionable; the trouble lies in their combinations. For example, confidence intervals for a single parameter are usually acceptable but, with many parameters, they are not. Even the ubiquitous sample mean for a normal distribution is unsound in high dimensions. In an experiment with several treatments, individual tests are fine but multiple comparisons present problems. Scientific truth is established by combining the results of many experiments: yet meta-analysis is a difficult area for statistics. How do you combine several data sets concerning the same hypothesis, each with its own significance level? The conclusions from two Student t-tests on means μ_1 and μ_2 do not cohere with the bivariate T-test for (μ_1, μ_2) (Healy, 1969). In contrast, the view adopted here easily takes the margins of the latter to provide the former.

The conclusion that probability is the only measure of uncertainty is therefore not just a pat on the back but strikes at many of the basic statistical activities. Savage developed his ideas in an attempt to justify statistical practice. He surprised himself by destroying some aspects of it. Let us therefore pass from disagreements to the constructive ideas that flow from the appreciation of the basic role of probability in statistics.

7. Inference

The formulation that has served statistics well throughout this century is based on the data having, for each value of a parameter, a probability distribution. This accords with the idea that the uncertainty in the data needs to be described probabilistically. It is desired to learn something

about the parameter from the data. Generally not every aspect of the parameter is of interest, so write it as (θ, α) where we wish to learn about θ with α as a nuisance, to use the technical term. Denoting the data by x, the formulation introduces $p(x|\theta, \alpha)$, the probability of x given θ and α. A simple example would have a normal distribution of mean θ and variance α, but the formulation embraces many complicated cases.

This handles the uncertainty in the data to everyone's satisfaction. The parameter is also uncertain. Indeed, it is that uncertainty that is the statistician's main concern. The recipe says that it also should be described by a probability $p(\theta, \alpha)$. In so doing, we depart from the conventional attitude. It is often said that the parameters are *assumed* to be random quantities. This is not so. It is the axioms that are assumed, from which the randomness property is deduced. With both probabilities available, the probability calculus can be invoked to evaluate the revised uncertainty in the light of the data:

$$p(\theta, \alpha|x) \propto p(x|\theta, \alpha)\, p(\theta, \alpha), \tag{1}$$

the constant of proportionality dependent only on x, not the parameters. Since α is not of interest, it can be eliminated, again by the probability calculus, to give

$$p(\theta|x) = \int p(\theta, \alpha|x)\, \mathrm{d}\alpha. \tag{2}$$

Equation (1) is the product rule; equation (2) the addition rule. Together they solve the problem of inference, or, better, they provide a framework for its solution. Equation (1) is Bayes's theorem and, by a historical accident, it has given its name to the whole approach, which is termed Bayesian. This perhaps unfortunate terminology is accompanied by some other which is even worse. $p(\theta)$ is often called the prior distribution, $p(\theta|x)$ the posterior. These are unfortunate because prior and posterior are relative terms, referring to the data. Today's posterior is tomorrow's prior. The terms are so engrained that their complete avoidance is almost impossible.

Let us summarize the position reached.

(a) Statistics is the study of uncertainty.
(b) Uncertainty should be measured by probability.
(c) Data uncertainty is so measured, conditional on the parameters.
(d) Parameter uncertainty is similarly measured by probability.
(e) Inference is performed within the probability calculus, mainly by equations (1) and (2).

Points (a) and (b) have been discussed and (c) is generally accepted. Point (e) follows from (a)–(d). The main protest against the Bayesian position has been to point (d). It is therefore considered next.

8. Subjectivity

At the basic level from which we started, it is clear that one person's uncertainties are ordinarily different from another's. There may be agreement over well-conducted games of chance, so that they may be used as a standard, but on many issues, even in science, there can be disagreement. As this is being written, scientists disagree on some issues concerning genetically modified food. It might therefore be sensible to reflect the subjectivity in the notation. The preferred way to do this is to include the concept of a person's knowledge at the time that the probability judgment is made. Denoting that by K, a better notation than $p(\theta)$ is $p(\theta|K)$, the probability of θ given K, changing to $p(\theta|x, K)$ on acquiring the data. This notation is valuable because it emphasizes the

fact that probability is always conditional (see Section 5). It depends on two arguments: the element whose uncertainty is being described and the knowledge on which that uncertainty is based. The omission of the conditioning argument often leads to confusion. The distinction between prior and posterior is better described by emphasizing the different conditions under which the probability of the parameter is being assessed.

It has been suggested that two people with the same knowledge should have the same uncertainties, and therefore the same probabilities. It is called the *necessary* view. There are two difficulties with this attitude. First, it is difficult to say what is meant by two people having the same knowledge, and also hard to realize in practice. The second point is that, if the probability does necessarily follow, it should be possible to evaluate it without reference to a person. So far no-one has managed the evaluation in an entirely satisfactory way. One way is through the concept of ignorance. If a state of knowledge I was identified, that described the lack of knowledge about θ, and that $p(\theta|I)$ was defined, then $p(\theta|K)$, for any K, could be calculated by Bayes's theorem (1) on updating I to K. Unfortunately attempts to do this ordinarily lead to a conflict with conglomerability. For example, suppose that θ is known only to assume positive integer values and you are otherwise ignorant about θ. Then the usual concept of ignorance means all values are equally probable: $p(\theta = i|I) = c$ for all i. The addition rule for n exclusive events $\{\theta = i\}$, with $nc > 1$, means $c = 0$ since no probability can exceed 1 by convexity. Now partition the positive integers into sets of three, each containing two odd, and one even, value: $A_n = (4n - 3,\ 4n - 1,\ 2n)$ will do. If E is the event that θ is even, $p(E|A_n) = \frac{1}{3}$ and by conglomerability $p(E) = \frac{1}{3}$. Another partition, each set with two even and one odd value, say $B_n = (4n - 2,\ 4n,\ 2n - 1)$, has $p(E|B_n) = \frac{2}{3}$ and hence $p(E) = \frac{2}{3}$ in contradiction with the previous result. By the suitable selection of a partition, $p(E)$ can assume any value in the unit interval. Such a distribution is said to be improper.

Unfortunately most attempts to produce $p(\theta|I)$ by a necessary argument lead to impropriety which, in addition to violating conglomerability, leads to other types of unsatisfactory behaviour. See, for example, Dawid *et al.* (1973). The necessary view was first examined in detail by Jeffreys (1961). Bernardo (1999) and others have made real progress but the issue is still unresolved. Here the view will be taken that probability is an expression by a person with specified knowledge about an uncertain quantity. The person will be referred to as 'you'. $p(A|B)$ is your belief about A when you know B. In this view, it is incorrect to refer to *the* probability; only to yours. It is as important to state the conditions as it is the uncertain event.

Returning to the example of θ taking positive integer values, as soon as you understand the meaning of the object to which the Greek letter refers, you are, because of that understanding, no longer ignorant. Then $p(\theta = i) = a_i$ with $\Sigma a_i = 1$. Some statistical research is vitiated by a confusion between the Greek letter and the reality that it represents. I challenge anyone to produce a real quantity about which they are truly ignorant. A further consideration is that a computer could not produce one example of θ. Since $p(\theta \leqslant n|I) = nc = 0$, $p(\theta > n) = 1$, so θ must surely be larger than any value that you care to name, or the computer can handle.

A further matter requires attention. It is common to use the same concept and notation when part of the conditioning event is unknown but assumed to be true. For example, all statisticians use $p(x|\theta, \alpha)$ as above, or, more accurately, $p(x|\theta, \alpha, K)$. Here they do not know the value of the parameter. What is being expressed is uncertainty about x, *if* the parameters were to have values θ and α (and they had knowledge K). It is not necessary to distinguish between supposition and fact in this context and the notation $p(A|B)$ is adequate.

The philosophical position is that your personal uncertainty is expressed through your probability of an uncertain quantity, given your state of knowledge, real or assumed. This is termed the subjective, or personal, attitude to probability.

Many people, especially in scientific matters, think that their statements are objective, expressed through *the* probability, and are alarmed by the intrusion of subjectivity. Their alarm can be alleviated by considering reality and how that reality is reflected in the probability calculus. We discuss in the context of science but the approach applies generally. Law provides another example. Suppose that θ is the scientific quantity of interest. (In criminal law, $\theta = 1$ or $\theta = 0$ according to whether the defendant did, or did not, commit the crime.) Initially the scientist will know little about θ, because the relevant knowledge base K is small, and two scientists will have different opinions, expressed through probabilities $p(\theta|K)$. Experiments will be conducted, data x obtained and their probabilities updated to $p(\theta|x, K)$ in the way already described. It can be demonstrated (see Edwards *et al.* (1963)) that under circumstances that typically obtain, as the amount of data increases, the disparate views will converge, typically to where θ is known, or at least determined with considerable precision. This is what is observed in practice: whereas initially scientists vary in their views and discuss, sometimes vigorously, among themselves, they eventually come to agreement. As someone said, the apparent objectivity is really a consensus. There is therefore good agreement here between scientific practice and the Bayesian paradigm. There are cases where almost all agree on a probability. These will be discussed when we consider exchangeability in Section 14.

It is now possible to see why there has been a reluctance to accept point (d), the use of a probability distribution to describe parameter uncertainty. It is because the essential subjectivity has not been recognized. With little data, $p(\theta, \alpha)$ varies among subjects: as the data increase, consensus is reached. Notice that $p(x|\theta, \alpha)$ is also subjective. This is openly recognized when two statisticians employ different models in their analysis of the same data set. We shall return to these points when the role of models is treated in Sections 9 and 11.

9. Models

The topic of models has been carefully discussed by Draper (1995) from the Bayesian viewpoint and that paper should be consulted for a more detailed account than that provided here. The philosophical position developed here is that uncertainty should be described solely in terms of your probability. The implementation of this idea requires the construction of probability distributions for all the uncertain elements in the reality being studied. The complete probability specification will be called a (probability) model, though the terminology differs from that ordinarily used in statistics, in a way to be described later. It also differs from model as used in science. The statistician's task is to construct a model for the uncertain world under study. Having done this, the probability calculus enables the specific aspects of interest to have their uncertainties computed on the knowledge that is available. There are therefore two aspects to our study: the construction of the model and the analysis of that model. The latter is essentially automatic; in principle it can be done on a machine. The former requires close contact with reality. To paraphrase and exaggerate de Finetti, *think* when constructing the model; with it, do not think but leave it to the computer. We repeat the point already made that, in doing this, the subject, whose probabilities are being sought, the 'you' in the language adopted here, is not the statistician, but the client, often a scientist who has asked for statistical advice. The statistician's task is to articulate the scientist's uncertainties in the language of probability, and then to compute with the numbers found. A model is merely your reflection of reality and, like probability, it describes neither you nor the world, but only a relationship between you and that world. It is unsound to refer to the true model. One time that this usage can be excused is when most people are agreed on the model. Thus the model of the heights of fathers and sons as bivariate normal might reasonably be described as true.

What uncertainties are there in a typical scenario? The fundamental problem of inference and induction is to use past data to predict future data. Extensive observations on the motions of heavenly bodies enables their future positions to be calculated. Clinical studies on a drug allow a doctor to give a prognosis for a patient for whom the drug is prescribed. Sometimes the uncertain data are in the past, not the future. A historian will use what evidence he has to assess what might have happened where records are missing. A court of criminal law enquires about what had happened on the basis of later evidence. We shall, however, use the temporal image, with past data x being used to infer future data y (as x comes before y in the alphabet). In this view, the task is to assess $p(y|x, K)$. In the interests of clarity, the background knowledge, fixed throughout the treatment, will be omitted from the notation and we write $p(y|x)$.

One possibility is to try to assess $p(y|x)$ directly. This is usually difficult, though it may be thought of as the basis of the apprenticeship system. Here an apprentice would sit at the master's feet and absorb the data x. With years of such experience, the apprentice could infer what would be likely to happen when he worked on his own. Successive observation on the use of ash in the construction of a wheel would enable him to employ ash for his own wheel. There is, however, a better way to proceed and that is to study the connections between x and y, and the mechanisms that operate. Newton's laws enable the tides to be calculated. Materials science assists in the design and construction of a wheel. Most modern inference can be expressed through a parameter θ that reflects the connection between the two sets of data. Extending the conversation, strictly within the probability calculus, to include θ,

$$p(y|x) = \int p(y|\theta, x)\, p(\theta|x)\, \mathrm{d}\theta.$$

It is usual to suppose that, once θ is known, the past data are irrelevant. In probability language, given θ, x and y are independent. Combining this fact with the determination of $p(\theta|x)$ by Bayes's theorem, we have

$$p(y|x) = \int p(y|\theta)\, p(x|\theta)\, p(\theta)\, \mathrm{d}\theta \Big/ \int p(x|\theta)\, p(\theta)\, \mathrm{d}\theta.$$

Now the task is to assess the uncertainty $p(\theta)$ about the parameter and the two data uncertainties, $p(x|\theta)$ and $p(y|\theta)$, given the parameter. Often the inference can stop at $p(\theta|x)$, leaving others to insert $p(y|\theta)$. This might happen with the drug example above, where the doctor would need to know the distribution of efficacy θ and then assess $p(y|\theta)$ for the individual patient.

Although much inference is rightly expressed in terms of the evaluation of $p(\theta|x)$, there is an important advantage in contemplating $p(y|x)$. The advantage accrues from the fact that y will eventually be observed; the doctor will see what happens to the patient. The parameter is not usually observed. The uncertainty of θ often remains; that of y disappears. This feature enables the effectiveness of the inference to be displayed by using a scoring rule in an extended version of that described in Section 4. If the inference is $p(y|x)$ and y is subsequently observed to be y_0, a score function $S\{y_0, p(\cdot|x)\}$ describes how good the inference was, so that the client and the statistician have their competences assessed. The method has been used in meteorology, e.g. in forecasting tomorrow's rainfall. Such methods are not readily available for $p(\theta|x)$. One of the criticisms that has been levelled against significance levels is that little study has been made of how many hypotheses, rejected at 5%, have subsequently been shown to be true. There is no reason to think that it is 5%. Theory suggests that it should be much higher and that significance is too easily attained. A weather forecaster who predicted rain on only 5% of days, when it subsequently rained on 20%, would not be highly esteemed. The best way to assess the quality of

inferences is to check $p(\theta|x)$ through the data probabilities $p(y|x)$ that they generate.

As previously mentioned, it is usually necessary to introduce nuisance parameters α, in addition to θ, to describe adequately the connection between x and y, and to establish the independence between them, given (θ, α). In the drug example, α might involve features of the individual patient. Nuisance parameters impose formidable problems for some forms of inference, like those based solely on likelihood, but, in principle, are easily handled within the probability framework by passing from the joint distribution of (θ, α) to the marginal for θ: equation (2).

The introduction of parameters reduces the construction of a model to providing $p(x|\theta, \alpha)$ and $p(\theta, \alpha)$. $p(y|\theta, \alpha)$ also arises but its assessment is similar to that for x and need not be separately discussed. Here we see the distinction between our use of 'model' and that commonly adopted, where only the data distribution, given the parameters, is included. Our definition includes the distribution of the parameters, since they form an important part of the uncertainty that is present. Most of the current literature on models therefore concerns the data and is discussed in the next section. For the moment, we just repeat the point made earlier that even $p(x|\theta, \alpha)$ is subjective. A common reason for wrongly thinking that it is objective lies in the fact that there is often more public information on the data than on the parameters, and we saw in Section 8 that, with increased information, people tend to approach agreement.

Why go through the ritual of determining $p(x|\theta, \alpha)$ and $p(\theta, \alpha)$, and then calculating $p(\theta|x)$? If $p(\theta, \alpha)$ can be assessed, why not assess $p(\theta|x)$ directly and avoid some complications? To use terminology that I do not like: if your prior can be assessed directly, why not your posterior? Part of the answer lies in the information that is typically available about the data density, but the desire for coherence is the major reason. A set of uncertainty statements is said to be *coherent* if they satisfy the rules of the probability calculus. Thus, the pair of statements $p(A|B) = 0.7$ and $p(A|\sim B) = 0.4$ do not cohere with the pair $p(B|A) = 0.5$ and $p(B|\sim A) = 0.3$. (Here $\sim B$ denotes the complement of B.) Think of A as a statement about data x and B as a statement about parameter θ. The first pair refers to uncertainties in the data and coheres with the first parameter statement, $p(B|A) = 0.5$, for data A. (Take $p(B) = 0.4/1.1 = 0.36$.) But all three do not cohere with the second parameter statement for data $\sim A$, that $p(B|\sim A) = 0.3$. With $p(B) = 0.36$, the coherent value is 0.22. The standard procedure ensures that you are prepared for any values of the data, A or $\sim A$, and the final inferences about B will collectively make sense.

10. Data analysis

Much statistical work is not concerned with a mathematical system, whether frequentist or Bayesian, but operates at a less sophisticated level. When faced with a new set of data, a statistician will 'play around' with them, an activity called (exploratory) data analysis. Elementary calculations will be made; simple graphs will be plotted. Several valuable ideas have been developed for 'playing', such as histograms and box plots. We argue that this is an essential, important and worthwhile activity that fits sensibly into the philosophy. The view adopted here is that data analysis assists in the formulation of a model and is an activity that precedes the formal probability calculations that are needed for inference. The argument developed so far in this paper has demonstrated the need for probability. Data analysis puts flesh onto this mathematical skeleton. The only novelties that we add to conventional data analysis is the recognition that its final conclusions should be in terms of probability and should embrace parameters as well as data. In the language of the last section, the conclusions of data analysis should cohere.

The fundamental concept behind the measurement of uncertainty was the comparison with a standard. Such comparisons are often difficult and there is a need to find some replacement. We do not measure length by using krypton light, the standard, but employ other methods. Data

analysis and the concept of coherence is such a replacement. Suppose that you need to assess a single probability; then all you have to guide you is the necessity that the value lies between 0 and 1. In contrast, suppose that the need is to assess several probabilities of related events or quantities, when the whole of the rich calculus of probabilities is available to help you in your assessments. In the example that concluded Section 9, you might have reached the four values given there, but considerations of coherence would force you to alter at least one of them. Coherence acts like geometry in the measurement of distance; it forces several measurements to obey the system. We have seen how this happens in replacing $p(y|x)$ by $p(x|\theta, \alpha)$ and $p(\theta, \alpha)$. Let us consider this and its relationship with data analysis, considering first the data density $p(x|\theta, \alpha)$.

A familiar and useful tool here is the histogram and modern variants like stem-and-leaf plots. These help to determine whether a normal density might be appropriate, or whether some richer family is required. If the data consist of two, or more, quantities $x = (w, z)$, then a plot of z against w will help to assess the regression of z on w and hence $p(z|w, \theta, \alpha)$. These devices involve the concept of repeated observations, e.g. to construct the histogram. We shall return to this point in discussion of the concept of exchangeability in Section 14.

There are issues here that have not always been recognized. You are making an uncertainty statement, $p(x|\theta, \alpha)$, for a quantity x, which, with the data available, is for you certain. Moreover you are doing it with real thought, in the form of data analysis about x. It is strange only to use uncertainty (probability) for the only certain quantity present. Furthermore, suppose that $t(x)$ describes the aspects of the data that you have considered, the histogram or the regression. Then the result of the data analysis is really $p\{x|\theta, \alpha, t(x)\}$; you are conditioning on $t(x)$. For example, you might say that x is normal with mean θ and variance α, but only after seeing $t(x)$, or equivalently doing the data analysis. This may lead to spurious precision in the subsequent calculations. One way to proceed would be to construct the model without looking at the data. Indeed, this is necessary when designing the experiment (Section 16). The construction could only come in close consultation with the client and would involve larger models than are currently used. Perhaps data analysis can be regarded as approximate inference, clearing out the grosser aspects of the larger model that are not needed in the operational, smaller model.

Another point is that $p(x|\theta, \alpha)$, say in the form of a histogram, is only exhibited for one value of (θ, α), namely the uncertain value that holds there. The data contain little evidence that, even if $x \sim N(\theta_0, \alpha_0)$, it is $N(\theta, \alpha)$ in situations unobserved. There is a case therefore for making models as big as your computing power will accommodate, to allow for non-normality and general parameter values. The size of a model is discussed in Section 11. Notice that the difficulties raised in the last two paragraphs are as relevant to the frequentist as they are to the Bayesian.

The assessment problem is different when it comes to the parameter density because there is often no repetition and the familiar tools of data analysis are no longer available. Furthermore, in handling the data density, several standard models are readily available, e.g. the exponential family and methods built around GLIM. These models have primarily been designed for ease of analysis through the possession of special properties like sufficient statistics of fixed low dimensionality, though they have the difficulty that outliers are not easily accommodated. These constraints have been imposed partly through limitations of computer capacity but more importantly because, within the frequency approach, there are no general principles and a new model may require the introduction of new ideas. Modern computational techniques lessen the first difficulty and Bayesian methods, with their ubiquitous use of the probability calculus, remove the second entirely; the object is always to calculate $p(\theta|x)$. We shall return to this point in Section 15.

Few standard models are available for the parameter density, essentially limited to the densities that are conjugate to the member of the exponential family chosen for the data density. The frequentist chant is 'where did you get that prior?'. It is not a silly gibe; there are serious

difficulties but they are partly caused by a failure to link theory and practice. I have often seen the stupid question posed 'what is an appropriate prior for the variance σ^2 of a normal (data) density?'. It is stupid because σ is just a Greek letter. To find the parameter density, it is essential to go beyond the alphabet and to investigate the reality behind σ^2. What is it the variance of? What range of values does the client think it has? Recall that the statistician's task is to express the uncertainty of you, the client, in probability terms. A sensible form of question might be 'what is your opinion about the variability of systolic blood pressure in healthy, middle-aged males in England?'. But, even with careful regard for practice, it would be stupid to deny the existence of very real, and largely unexplored, problems here. This is especially true when, as in most cases, the parameter space has high dimensionality. We are lacking in methods of appreciating multivariate densities. (This is true of data as well as parameters.) Physicists did not deny Newton's laws because several of the ideas that he introduced were difficult to measure. No, they said that the laws made sense, they work where we can measure, so let us develop better methods of measurement. Similar considerations apply to probability. A neglected area of statistical research is the expression of multivariate opinion in terms of probability, where independence is invoked too often, on grounds of simplicity, ignoring reality. It is not often recognized that the notion of independence, since it involves probability, is also conditional. The mantra that $(x_1, x_2, \ldots, x_n)$—forming a random sample—are independent is ridiculous when they are used to infer x_{n+1}. They are independent, given θ.

It is sometimes argued that data analysis can make no contribution to the assessment of a distribution for the parameter because it involves looking at the data, whereas what is needed is a distribution prior to the data. This is countered by the observation that we all use data to suggest something and then consider what our attitude to it was without the data. You see a sequence of 0s and 1s and notice few, but long, runs. Could the sequence be Markov instead of exchangeable as you had anticipated? You think about reasons for the dependence and, having decided that a Markov chain is possible, think about its value. Had you seen 1s only when the order was prime, you would fail to find reasons and accept the extraordinary thing that has happened.

11. Models again

A model is a probabilistic description of a client's situation, whose assessment is helped by data analysis and exploration of the client's present understanding. Several problems remain, of which one is the size of the model. Should you include extra quantities, besides x, as covariates? Should the parameters increase in number to offer greater flexibility, replacing a normal distribution by a Student's t, say? Savage once gave the wise advice that a model should be as big as an elephant. Indeed, the ideal Bayesian has one model embracing everything: what has been termed a world view. Such a model is impractical and you must be content with a small world embracing your immediate interests. But how small should it be? Really small worlds have the advantage of simplicity and the possibility of obtaining many results, but they have the disadvantage that they may not capture your understanding of reality so that $p(y|x)$ based on them may have a high penalty score. Compromise is called for, but always choose the largest model that your computational powers will tolerate. One successful strategy is to use a large model and to determine, through robustness studies, what aspects of the model seriously affect your final conclusion. Those that do not can be ignored and some reduction in size achieved.

It is valuable to think about the relationships between the small world selected and the larger worlds that contain it. In England it is current practice to publish league tables of schools' performances that use only examination results. Many contend that this is a ridiculously small world and that other quantities, like the performance of pupils at admission, should be included. It

is sometimes said that, in our approach, a smaller world cannot fit into a larger one and that, if the former is found to be inadequate, it is necessary to start afresh. This is not so; the apprehension arises through a failure to appreciate the conditional nature of probability. Here is an example. Suppose that your model is that $x \sim N(\theta, \alpha)$. Then, in full, you are describing $p(x|\theta, \alpha, N, K)$, where N denotes normality. In words, knowing K and supposing normality, x has mean θ and variance α. If the presence of outliers suggests an extension to Student's t, so that $x \sim t(\theta, \alpha, \nu)$ with index ν, then the two models cohere, the former having the restriction, or condition, that $\nu = \infty$. In contemplating t, you will already have considered large values of ν. Typically, in passing from a small to a large model, the former will correspond to the latter under conditions and the smaller is embedded in the larger. Occasionally, this appears not to be so. For example, one model may say that x is normal; the other that $\log(x)$ is. One way out of this difficulty is to introduce a series of transformations with x and $\log(x)$ as two members, as suggested by Box and Cox (1964). If this is not possible and you are genuinely uncertain whether model M_1 or model M_2 obtains, then describe your uncertainty by probability, producing a model that has M_1 with probability γ and M_2 with probability $1 - \gamma$. Part of the inferential problem will be the passage from γ to $p(M_1|x)$. This is a problem that has been discussed (O'Hagan, 1995), and where impropriety is best avoided and conglomerability assumed.

Large models have been criticized because they can sometimes appear to produce unsatisfactory results in comparison with smaller models. For example, in considering the regression of one quantity on many others, you are urged not to include too many regressor variables, because to do so leads to overfitting. This undesirable feature comes about through the use of frequentist methods. A theorem within the Bayesian paradigm shows that the phenomenon cannot arise with a coherent analysis, essentially because maximization over a subset cannot exceed that over the full set. The issue is connected with conglomerability (Section 5) because the method of fitting that is ordinarily used, least squares, is equivalent to a Bayesian argument using an improper prior, namely a uniform distribution over the space of the regression parameters. This does not cause offence when the dimension of the space is low, but causes increasing difficulties as it grows (Stein, 1956), and hence the overfitting.

Statisticians have, over the years, developed a collection of standard models, some of which are so routine that computer packages for their implementation exist. Although these, when modified from their frequentist form to provide a coherent analysis, are indubitably valuable, they should never replace your careful construction of a model from the practical realities. We repeat the important advice to *think* in constructing the model: once that has been done, leave everything to the probability calculus. An illustration is provided by the inconvenient phenomenon of non-response in sample surveys. Here it is important to think about the mechanisms that gave rise to the lack of response, and to model them. Some models in the literature do not flow from any real understanding of why the data are incomplete, and they are therefore suspect. The client's reality must be modelled in probability terms.

The suggestion has often been made that it is possible to test the adequacy of a model, without the specification of alternatives, and methods for doing this have been developed (Box, 1980). We argue that the rejection of a model is not a reality except in comparison with an alternative that appears better. The reason lies in the nature of probability which is essentially a comparative measure. The Bayesian world is a comparative world in which there are no absolutes. The point will emerge again in decision analysis in Section 15, where you decide to do something, not because it is good, but because it is better than anything else that you can think of. People who refuse to vote in an election on the grounds that no candidate meets their requirements miss the point that, with the limited availability of candidates, you should choose the one whom you think is best, even if awful.

12. Optimality

The position has been reached that the practical uncertainties should be described by probabilities, incorporated into your model and then manipulated according to the rules of the probability calculus. We now consider the implications that the manipulations within that calculus have on statistical methods, especially in contrast with frequentist procedures, thereby extending the discussion of significance tests and confidence intervals in Section 6. It is sometimes said, by those who use Bayes estimates or tests, that all the Bayesian approach does is to add a prior to the frequentist paradigm. A prior is introduced merely as a device for constructing a procedure, that is then investigated within the frequentist framework, ignoring the ladder of the prior by which the procedure was discovered. This is untrue: the adoption of the full Bayesian paradigm entails a drastic change in the way that you think about statistical methods.

A large amount of effort has been put into the derivation of optimum tests and estimates. This is evident on the theoretical side where the splendid scholarly books of Lehmann (1983, 1986) are largely devoted to methods of finding good estimates and tests respectively. Again, more informally, in data analysis, reasons are advanced for using one procedure rather than another, as when trimmed means are rightly said to be better than raw means in the presence of outliers. Let us therefore look at inference, in the sense of saying something about a parameter θ, given data x, in the presence of nuisance parameters α. The frequentist may seek the best point estimate, confidence interval or significance test for θ.

A remarkable, and largely unrecognized, fact is that, within the Bayesian paradigm, all the optimality problems vanish; a whole industry disappears. How can this be? Consider the recipe. It is to calculate $p(\theta|x, K)$, the density for the parameter of interest given the data and background knowledge. This density is a complete description of your current understanding of θ. There is nothing more to be said. It is an estimate: your only estimate. Integrated over a set H, it provides your entire understanding of whether H is true. There is nothing better than $p(\theta|x, K)$. It is unique; the only candidate. Consider the case of the trimmed means just mentioned. If the model incorporates simple normality, the density for θ is approximately normal about $\bar{x}$, the sample mean. However, suppose that normality is replaced by Student's t (with two nuisance parameters, spread and degrees of freedom); then the density for θ will be centred, not on $\bar{x}$, but on what is essentially a trimmed mean. In other words, the estimate arises inevitably and not because of optimality considerations.

The Bayesian's unique estimate, the posterior distribution, depends on the prior, so there is some similarity between the Bayesian and the frequentist who uses a prior to construct their optimum estimates. (The class of good frequentist procedures is the Bayes class.) The real difference is that the frequentist will use different criteria, like the error rate, rather than coherence to judge the quality of the resulting procedure. This is discussed further in Section 16.

13. The likelihood principle

We have seen that parametric inference is made by calculating

$$p(\theta|x) = p(x|\theta)\, p(\theta) \Big/ \int p(x|\theta)\, p(\theta)\, \mathrm{d}\theta. \tag{3}$$

Consider $p(x|\theta)$ as a function of two quantities, x and θ. As a function of x, for any fixed θ, $p(\cdot|\theta)$ is a probability density, namely it is positive and integrates, over x, to 1. As a function of θ, for any fixed x, $p(x|\cdot)$ is positive but does not usually integrate to 1. It is called the likelihood of θ for the fixed x. It is immediate from equation (3) that the only contribution that the data make to inference is through the likelihood function for the observed x. This is the likelihood principle that

values of x, other than that observed, play no role in inference. A valuable reference is Berger and Wolpert (1988).

This fact has important recognized consequences. Whenever in inference an integration takes place over values of x, the principle is violated and the resulting procedure may cease to be coherent. Unbiased estimates and tail area significance tests are among the casualties. The likelihood function therefore plays a more important role in Bayesian statistics than it does in the frequentist form, yet likelihood alone is not adequate for inference but needs to be tempered by the parameter distribution. Uncertainty must be described by probability, not likelihood. Before enlarging on this remark, it is important to be clear what is meant by likelihood. If a model with data x has been developed with parameters (θ, α), then $p(x|\theta, \alpha)$ as a function of (θ, α), for the fixed observed value of x, is undoubtedly the likelihood function. However, inference in equation (3) does not involve the entire likelihood function, but only its integral

$$p(x|\theta) = \int p(x|\theta, \alpha)\, p(\alpha|\theta)\, \mathrm{d}\alpha. \tag{4}$$

We refer to this as the likelihood of θ but the terminology is not always accepted. The reason is clear: its construction involves one aspect, $p(\alpha|\theta)$, of the parameter density, $p(\theta, \alpha)$, which latter is not admitted to the frequentist or likelihood schools. In neither school is there general agreement about what constitutes the likelihood function for a parameter θ of interest in the presence of a nuisance parameter α. There are at least a dozen candidates in the literature. For example, in addition to the integrated form in equation (4), there is $p(x|\theta, \hat{a})$, where $\hat{a}$ is the value that makes $p(x|\theta, \alpha)$ over α a maximum. The plethora of candidates reflects the impossibility of any satisfactory definition that avoids the intrusion of probabilities for the parameters.

The reason for likelihood being, on its own, inadequate is that, unlike probability, it is not additive. If A and B are two exclusive sets, then $p(A \cup B) = p(A) + p(B)$, omitting the conditions, whereas it is not true that $l(A \cup B) = l(A) + l(B)$ for a likelihood function $l(\cdot)$. Since the properties used as axioms in the development of inference, e.g. in the work of Savage, lead to additivity, any violation may lead to some violation of the axioms. This happens with likelihood. In Section 5 we had an example involving colour and sex, which was expressed in terms of the informal concept of one event being more likely than another. In fact, the example holds when 'likely' is used in the technical sense as defined here. Likelihood is an essential ingredient in the inference recipe but it cannot be the only one.

Notice that the likelihood principle only applies to inference, i.e. to calculations once the data have been observed. Before then, e.g. in some aspects of model choice, in the design of experiments or in decision analysis generally, a consideration of several possible data values is essential (see Section 16).

14. Frequentist concepts

Ever since the 1920s, statistics has been dominated by the frequentist approach and has, by any sensible criterion, been successful; yet we have seen that it clashes with the coherent view in apparently serious ways. How can this be? Our explanation is that there is a property, shared by both views, that links them more closely than the material so far presented here might suggest. The link is the concept of exchangeability. A sequence $(x_1, x_2, \ldots, x_n)$ of uncertain quantities is, for you, exchangeable under conditions K if your joint probability distribution, given K, is invariant under a permutation of the suffixes. For example, $p(x_1 = 3, x_2 = 5|K) = p(x_2 = 3, x_1 = 5|K)$ on permuting 1 and 2. An infinite sequence is exchangeable if every finite subsequence is so judged. The roles of 'you' and K have been mentioned to emphasize that exchangeability is a

subjective judgment and that you may change your opinion if the conditions change.

If you judge a sequence to be (infinitely) exchangeable, then your probability structure for the sequence is equivalent to introducing a parameter, ψ say, such that, given ψ, the members of the sequence are independent and identically distributed (IID). As the parameter is uncertain, you will have a probability distribution for it. This result is due to de Finetti (1974, 1975). Ordinarily ψ will consist of elements (θ, α) of which θ is of interest and α is nuisance. Consequently exchangeability imposes the structure used above but with the addition that the data x have the particular form of IID components. Furthermore, ψ is related to frequency properties of the sequence. Thus, in the simple case where x_i is either 0 or 1, the Bernoulli sequence, ψ is the limiting proportion of them that are 1. Consequently, a Bayesian who makes the exchangeability judgment is effectively making the same judgment about data as a frequentist, but with the addition of a probability specification for the parameter.

The concept of IID observations has dominated statistics in this century. Even when obviously inappropriate, as in the study of time series, the modelling uses IID as a basis. For example, $x_n - \theta x_{n-1}$ may be supposed IID for some θ, leading to a linear, autoregressive, first-order process. Within the IID assumption, frequency ideas are apposite, some even within the Bayesian canon, so there has developed a belief that uncertainty and probability are therefore based on frequency. Some statistics texts only deal with IID data and therefore restrict the range of statistical activities. Their examples will come from experimental science, where repetition is basic, and not from law, where it is not. Frequency, however, is not adequate because there is ordinarily no repetition of parameters; they have unique unknown values. Consequently the confusion between frequency and probability has denied the frequentist the opportunity of using probability for parameter uncertainty, with the result that it has been necessary for them to develop incoherent concepts like confidence intervals.

The use of frequency concepts outside exchangeability leads to another difficulty. Frequentists often support their arguments by saying that they are justified 'in the long run', to which the coherent response is 'what long run?'. For example, a confidence interval (see Section 6) will cover the true value a proportion $1 - \alpha$ of times in the long run. To make sense of this it is necessary to embed the particular case of data x into a sequence of similar data sets: which sequence?; what is similar? The classic example is a data set consisting of r successes in n trials, judged to be Bernoulli. In the sequence do we fix n, or fix r or some other feature of the observed data? It matters. Bayesians provide an answer for the single situation, whereas frequentists often need to embed the situation into a sequence of situations.

The restriction of probability to frequency can lead to misrepresentations. Here is an example, concerning the determination of physical constants, such as the gravitational constant G. It is common and reasonable to suppose that the measurements made at one place and time are exchangeable and unbiased, each having expectation G. It is reasonable to use their mean as the current estimate of G. Some rejection of outliers may be needed before this is done. The uncertainty attached to this estimate is found by taking s^2, equal to the average of the squared deviations from the mean, and quoting a standard error of $s/\sqrt{n}$, where n is the number of measurements. This leads to confidence limits for G. Experience shows that the more recent estimates usually lie outside the confidence limits of earlier estimates. In other words, the limits were too narrow. A scoring rule for estimators of G would produce a large penalty score. The reason is that the measurements are actually biased. Since the amount of the bias is not amenable to frequency ideas, it is ignored. The Bayesian approach would have a distribution for the bias and would use as a prior for G the posterior from the last estimate, possibly adjusted for any modifications in the measurement process. Often standard errors are too small because only the exchangeable component of uncertainty is considered. Similar mistakes can arise with the

predictions of future numbers of cases of acquired immune deficiency syndrome. They can ignore changes in personal behaviour or Government policy, changes that are not amenable to frequentist analysis.

15. Decision analysis

It has been noted how statistics began with the collection and presentation of data, and then extended to include the treatment of the data and the process which we now call inference. There is a further stage beyond that, namely the use of data, and the inferences drawn from them, to reach a decision and to initiate action. In my view, statisticians have a real contribution to make to decision analysis and should extend their data collection and inference to include action. The methods of Ramsey and Savage have demonstrated how the foundations can be presented through decision analysis. The extension to include action can be better understood if we ask what is the purpose of an inference that consists in calculating $p(y|x)$ for future data y, conditional on past data x. An example cited was a doctor who had data on a drug and wished to infer what might happen to a patient given the drug. The example involves a decision, namely which drug to prescribe, another drug possibly leading to a different inference for y. We argue, following Ramsey, that an inference is only of value if it is capable of being used to initiate action. Partial knowledge that cannot be used is of little value. Even in its parametric form, $p(\theta|x)$ will only be worthwhile if it can be incorporated into actions that involve the uncertain θ. Marx was right: the point is not just to understand the world (inference) but also to change it (action). Let us see how this can be done in the Bayesian view.

The structure used by Savage and others is to formulate a list of possible decisions d that might be taken. The uncertainty is captured in a quantity (or parameter) θ. The pair (d, θ) is termed a consequence, describing what will happen if you take decision d when the parameter has value θ. We have seen how the uncertainty in θ needs to be described by a probability distribution $p(\theta)$. This will be conditional on your state of knowledge, which is omitted from the notation. It may also depend on the decision, as in the case where the decisions are to invest in advertising, or not, and θ is next year's sales. We therefore write $p(\theta|d)$. The foundational argument goes on to show that the merits of the consequence (d, θ) can be described by a real number $u(d, \theta)$, termed the utility of the consequence. One consequence is preferred to another if it has the higher utility. If these utilities are constructed in a sensible way, the best decision is that which maximizes your expected utility

$$\int u(d, \theta)\, p(\theta|d)\, \mathrm{d}\theta.$$

The addition of a utility function for consequences, combined with the probability description of uncertainty, leads to a solution to the decision problem. Utility has to be described with care. It is not merely a measure of worth, but a measure of worth on a probability scale. If the best consequence has utility 1 and worst utility 0, then consequence (d, θ) has utility $u(d, \theta)$ if you (notice the subjective element) are indifferent between

(a) the consequence for sure and
(b) a chance $u(d, \theta)$ of the best (and $1 - u(d, \theta)$ of the worst).

It is this probability construction that enables the expectation to emerge as the only relevant criterion for the choice of decision. Utility embraces all aspects of the consequence. For example, if one outcome of a gamble is a win of £100, its utility includes not only an increase in monetary assets but also the thrill of the gamble. Some analyses, based solely on money, are defective because of their limited view of utility.

Notice that, just as $p(\theta)$ is not the statistician's uncertainty, but rather the client's, so the utility is that of the decision maker. The statistician's role is to articulate the client's preferences in the form of a utility function, just as it is to express their uncertainty through probability. Notice also that the analysis supposes that there is only one decision maker, the 'you' of our text, though 'you' may be several individuals forming a group, making a collective decision. None of the arguments given here apply to the case of two, or more, decision makers who do not have a common purpose, or may even be in conflict. This is an important limitation on maximized expected utility.

One topic that statisticians have often considered their own, at least since the brilliant work of Fisher (1935), is the design of experiments. This is a decision problem and fits neatly into the principles just enunciated. Let e be a member of a class of possible experiments from which one must be selected. Let x denote data that might arise from such an experiment. The experimentation presumably has some purpose, expressed by the selection of a (terminal) decision d. As usual, denote the uncertain element by θ. (A similar analysis applies when the inference is for future data y.) The final consequence of experimentation and action is (e, x, d, θ) to which you attach a utility $u(e, x, d, \theta)$. The expected utility is

$$\int u(e, x, d, \theta)\, p(\theta|e, x, d)\, \mathrm{d}\theta, \tag{5}$$

the uncertainty being conditional on all the other ingredients. The optimum decision is that d which maximizes expression (5). Denote the maximum value so obtained by $\overline{u}(e, x)$. The expectation of this is

$$\int \overline{u}(e, x)\, p(x|e)\, \mathrm{d}x, \tag{6}$$

since x is the only uncertain element at this stage, the uncertainty being clearly dependent on the experiment e. A final maximization of expression (6) provides the optimum experimental design.

Notice the simplicity of the principles that are involved here, even though the technical manipulations may be formidable. There is a temporal sequence that alternates between taking an expectation over the quantities that are uncertain and maximizing over the decisions that are available. Each uncertainty must be evaluated conditionally on all that is known then. The utility is attached to the final outcome, other (expected) utilities, like $\overline{u}(e, x)$ being derived therefrom. This provides a formal framework for the design of experiments.

It would appear to be a sensible criticism of the method just outlined that many experiments are not conducted with a terminal decision in mind but merely to gather information about θ. This aspect can be accommodated by extending the interpretation of a decision. Information about θ depends on your uncertainty about θ expressed, as always, by probability. So let the decision d be to select the relevant density, here $p(\theta|e, x)$. A utility function can then be constructed. Often it is reasonable to suppose u additive in the sense that

$$u(e, x, d, \theta) = u(e, x) + u(d, \theta), \tag{7}$$

the first term involving the experimental cost and the second the terminal consequences. Notice the connection between $u(d, \theta)$ and the scoring rules suggested in Section 9. Here $u(d, \theta_0)$ may be thought of as a reward score attached to decision d to announce $p(\theta|e, x)$ when the parameter has value θ_0. The usual measure of the information provided by $p(\theta)$ is Shannon's,

$$\int p(\theta) \log\{p(\theta)\}\, \mathrm{d}\theta.$$

The language of decision analysis has been used by Neyman and others in connection with

hypothesis testing, where they speak of the decisions to accept, and to reject, the hypothesis H. There are cases where acceptance and rejection can legitimately be thought of as action, as with the rejection of a batch of items. Equally there are other cases where we could calculate $p(H|x, K)$ as an inference about H on data x and knowledge K. The latter form may, as in the last paragraph, be thought of as a decision. Both forms are valid and useful for different purposes. Our philosophy accommodates both views and it is for you to consider how to model the reality before you. An important feature of the Bayesian paradigm is its ability to encompass a wide variety of situations using a few basic principles.

Some writers, in discussing hypothesis testing, have argued that there are many different cases. For example, some may really involve action; some are purely inferential. Other cases have been described, ending up, as with likelihood, in a plethora of situations and great complexity. The Bayesian view is that these are all covered by the general principles and that the differences perceived are differences in the probability and utility structures. Some folk love complexity for it hides inadequacies and even errors.

16. Likelihood principle (again)

In Section 13 it was seen how the likelihood principle is basic for inference, yet denied by many frequentist notions. The principle ceases to apply when experimental design is part of the decision analysis, essentially because of the integration over x involved in expression (6). At the initial stage, where you are considering which experiment to perform, the data, conditional on any experiment selected, is uncertain for you. This uncertainty is expressed through $p(x|e)$ and is eliminated by the operation of expectation in expression (6). In conducting an inference, or in making a terminal decision, you know the value of x, for the data are available. Consequently it is unnecessary to consider other data values and the likelihood is all that is needed. When it is a question of experimental design, the data are surely not available and all possibilities must be contemplated. This contrast between pre- and post-data emphasizes the importance of the conditions when you face uncertainty. Probability is a function of *two* arguments, not one.

Just how the consideration of the experiment can involve one form of integration over x used by frequentists, namely error rates, can be seen as follows. Denote by $d^*(e, x)$ that decision which maximizes the expected utility (5). The expectation over x, expression (5), can then be written

$$\int\int u\{e, x, d^*(e, x), \theta\}\, p\{\theta|e, x, d^*(e, x)\}\, \mathrm{d}\theta\, p(x|e)\, \mathrm{d}x.$$

Now

$$p\{\theta|e, x, d^*(e, x)\} = p(\theta|e, x)$$

since the addition of d^*, a function of e and x, adds no further condition. The latter probability is

$$p(x|e, \theta)\, p(\theta|e)/p(x|e).$$

Inserting this value into the expectation and reversing the orders of integration, we have

$$\int\int u\{e, x, d^*(e, x), \theta\}\, p(x|e, \theta)\, \mathrm{d}x\, p(\theta|e)\, \mathrm{d}\theta$$

where the inner integral exposes the frequentist integration over x. For a fixed experiment, and with a utility that does not directly involve x, the relevant integral is

$$\int u\{d^*(e, x), \theta\}\, p(x|e, \theta)\, \mathrm{d}x.$$

With two decisions and 0–1 utility, we immediately have $\int p(x|e, \theta)\,\mathrm{d}x$ over a subset of sample space and the familiar errors of the two kinds.

The occurrence of error rates leads to some confusion because they are often treated as the quantities to be controlled, and therefore occupy a primary position in decision analysis, whereas our primary consideration lies in the utility structure. Once the utility structure has been imposed, the errors will look after themselves. However, a consideration of different errors may lead to undesirable changes in the utility structure. The Bayesian view is that the utilities, not the errors, are the invariants of the analysis. For example, to design an experiment to achieve prescribed error rates may be incoherent. The prescription should instead specify utilities.

17. Risk

Risk is a term which we have not used. It has been defined (Duckworth, 1998) as 'the potential for exposure to uncertain events, the occurrence of which would have undesirable consequences'. The definition recognizes the two elements in what we have called decision analysis, the uncertainty and the utility, though Duckworth, in common with most statisticians, emphasized the loss, or undesirability, rather than the gain, the utility. The change is linguistic. Risk is therefore dependent on two arguments and our foundational presentation in Section 3 is dependent on the separation of them in uncertainty and worth. Yet it is common, as Duckworth does, to quote a measure of risk as a single number, so denying the separation. Thus the risk associated with a 1000-mile flight is 1.7 in suitable units. This is defensible for the following reason.

The optimum decision maximizes expected utility which, for data x, is proportional to

$$\int u(d, \theta)\, p(\theta)\, p(x|\theta)\,\mathrm{d}\theta$$

and may be written as a weighted likelihood

$$\int w(d, \theta)\, p(x|\theta)\,\mathrm{d}\theta,$$

where $w(d, \theta) = u(d, \theta)\, p(\theta)$. The analysis, for given data, and hence for a given likelihood, does not depend separately on the utility and probability, the two corner-stones of the philosophy, but only on their product. To put it in another way, if you were to watch a coherent person acting (as distinct from expressing his thoughts) you would not, on the basis of the observed actions, be able to separate the two elements; only the weight function might be determined.

Nevertheless there are several reasons for separating utility from probability. The most important is the need for inference, i.e. for a sound appreciation of the world without reference to action. The philosophy says that this is had through your probability structure for the world. In inference, manipulations take place entirely within the probability calculus, which therefore becomes separated from utility. There are people who argue that inference, in the form of pure science, is unsatisfactory when isolated from its applications in technology. What is undoubtedly important is that inference should be in a suitable form for decision-making, not an activity that is isolated from application. We have seen how Bayesian inference is perfectly adapted for this purpose. It will be seen in Section 19 how some aspects of the law separate inference from decision.

Another reason for the separation lies in the desirability of communication between people, between different 'yous'. Take the example of the 1000-mile flight cited above. Part of the calculation rests on the observed accident rate for aircraft. Another part rests on the consequences of the flight. You may react differently to these two elements. For example, it is known that for

elderly people there is an increased risk of circulatory problems due to sitting for hours in cramped seats, and therefore you may evaluate your accident rate differently from that suggested purely by the accident rate for aircraft. In contrast, a healthy, middle-aged executive, travelling in more comfort in first class, may accept the accident statistics but have a different utility because of the importance of the meeting to which he is bound. These considerations suggest that the accident rate and consequences of an accident be kept separate because you may be able to use one element but not the other, whereas the weight function alone would be more difficult to use.

18. Science

Karl Pearson said 'The unity of all science consists alone in its method, not in its material' (Pearson, 1892). It is not true to say that physics is science whereas literature is not. There are times when a physicist makes a leap of the imagination like an artist. Analyses of word counts can help to identify the author of an anonymous piece of literature. Scientific method is certainly much more important in physics than in literature, but it has the potentiality to be used in any discipline.

Of what then does the method consist? There is an enormous literature devoted to answering this question and it is presumptuous of me to claim to have the answer. But I do believe that statisticians, in their deep study of the collection and analyses of data have, perhaps unwittingly, uncovered the answer and it lies in the philosophy presented here. Experimentation, with its production of data, is an essential ingredient of scientific method, so the connection between statistics and science is not surprising. In this view, the scientific method consists in expressing your view of your uncertain world in terms of probability, performing experiments to obtain data, and using that data to update your probability and hence your view of the world. Although the emphasis in this updating is ordinarily put on Bayes, effectively the product rule, the elimination of the ubiquitous nuisance parameters by the addition rule 2 is also important. As we have seen, the design of the experiment is also amenable to statistical treatment. Scientific method consists of a sequence alternating between reasoning and experimentation. As explained in Section 8, each scientist is a 'you' with their own beliefs which are brought into harmony through the accumulation of data. It is this consensus that is objective science.

Objections have been made to this simple view on the grounds that scientists do not act in the way described in the last paragraph. They even do tail area significance tests. The response to the objection is that our philosophy is normative, not descriptive. It is not the intention to describe how scientists behave but how they would wish to behave if only they knew how. The probability calculus provides the 'how'. An impediment affecting 'how' is the lack of good methods of assessing probabilities when no exchangeability assumption is available to guide you. This is ordinarily described as determining your prior, but in reality it is wider than that. Some attacks on science are truly attacks on how scientists behave—on the descriptive aspect. Often they are valid. Such attacks would become less cogent if they dealt with the normative aspect. Scientists are human. Real scientists are affected by extraneous conditions. One would hope that a scientist working for a multinational company and another employed by an environmental agency differ only in their probabilities and would update accordingly. One suspects that other issues intervene. It is my hope that a Bayesian approach would help to expose any biases or fallacies in either of the protagonists' arguments.

19. Criminal law

There are two reasons for including this section on criminal law: first because of my own interest in forensic science; second because of the conviction that this interest has engendered that some

important aspects of the law are amenable to the scientific method as described in the last section. These aspects concern the trial process, where there is uncertainty about the defendant's guilt, uncertainty that is subsequently tempered by data, in the form of evidence, hopefully to reach a consensus about the guilt. Clearly this fits into the paradigm developed here. Lawyers do not have a monopoly on the discovery of the truth; scientists have been doing it successfully for centuries. There are aspects of the law, like the writing of a law, to which the scientific method has little to contribute. However, the courts are not just concerned with guilt; they need to pass sentence. The law has separated these two functions, just as we have. They can be recognized as inference and decision-making respectively.

The defendant in a court of law is either truly guilty G or not guilty $\sim G$. The guilt is uncertain and so should be described by a probability $p(G)$. (The background knowledge is omitted from the notation.) Data, in the form of evidence E, are produced and the probability updated. Since there are only two possibilities, G or $\sim G$, it is convenient to work in terms of odds (on),

$$o(G) = p(G)/p(\sim G),$$

when Bayes's theorem reads

$$o(G|E) = \frac{p(E|G)}{p(E|\sim G)} o(G)$$

involving multiplication of the original odds by the likelihood ratio. Evidence often involves nuisance parameters but, in principle, these can be eliminated in the usual way by the addition rule. They will often enter into $p(E|\sim G)$ because there might be several ways in which the crime could have been committed, other than by the defendant. As the trial proceeds, further evidence is introduced and successive multiplications by likelihood ratios determine the final odds. A difficulty here is that successive pieces of evidence may not be independent, either given G or given $\sim G$.

So far this method has mainly been used successfully for scientific evidence, like bloodstains and DNA (Aitken and Stoney, 1991). Its applicability in general depends on satisfactory methods of probability assessment. It has the potential advantage of helping the court to combine disparate types of evidence for, as remarked in Section 3, the principal merit of measurement lies in its ability to meld several uncertainties into one.

The law agrees with the philosophy in separating inference from decision. It even allows different evidence to be admitted into the two processes. For example, previous convictions may be used in sentencing (decision) but not always in assessing guilt (inference). Expected utility analysis includes a theorem to the effect that cost-free information is always expected to increase the utility. This suggests that the only reason for not admitting evidence should be on grounds of cost (Eggleston, 1983).

The part of the trial process that, at present, results in the judgment guilty, or not guilty, should, in our view, be replaced by the calculation of odds $o(G|E)$, where E is now the totality of all admitted evidence. On this view, the jury should not make a firm statement of guilt, or not, but state their final odds, or probability, of guilt. At least this provides a more flexible and informative communication. More importantly, it provides the judge with the information that he needs for sentencing. If d is a possible decision, about gaol or a fine, then the expected utility of d is

$$u(d, G)\, p(G|E) + u(d, \sim G)\, p(\sim G|E).$$

The optimum sentence is that d which maximizes this expectation. The utilities here will reflect society's evaluation of the merits of different sentences for the guilty person, and the seriousness

of false imprisonment. We are a long way from the implementation of these ideas but even now they can guide us into sensible procedures and avoid incoherent ones.

20. Conclusions

The philosophy of statistics presented here has three fundamental tenets: first, that uncertainty should be described by probabilities; second, that consequences should have their merits described by utilities; third, that the optimum decision combines the probabilities and utilities by calculating expected utility and then maximizing that. If these are accepted, then the first task of a statistician is to develop a (probability) model to embrace the client's interests and uncertainties. It will include the data and any parameters that are judged necessary. Once accomplished, the mechanics of the calculus take over and the required inference is made. If decisions are involved, the model needs to be extended to include utilities, followed by another mechanical operation of maximizing expected utility. One attractive feature is that the whole procedure is well defined and there is little need for *ad hoc* assumptions. There is, however, a considerable need for approximation. To carry out this scheme for the large world is impossible. It is essential to use a small world, which introduces simplification but often causes distortion. Even the mechanics of calculation need numerical approximations. Both these issues have been considered in the literature, whether frequentist or Bayesian, and substantial progress has been made. Where a real difficulty arises is in the construction of the model. Many valuable techniques have been introduced but, because of the frequentist emphasis in past work, there is a real gap in our appreciation of how to assess probabilities—of how to express our uncertainties in the requisite form. My view is that the most important statistical research topic as we enter the new millennium is the development of sensible methods of probability assessment. This will require co-operation with numerate experimental psychologists and much experimental work. A colleague put it neatly, though with some exaggeration: 'There are no problems left in statistics except the assessment of probability'. It is curious that the typical expert in probability knows nothing about, and has no interest in, assessment.

The adoption of the position outlined in this paper would result in a widening of the statistician's remit to include decision-making, as well as data collection, model construction and inference. Yet it also involves a restriction in their activity that has not been adequately recognized. Statisticians are not masters in their own house. Their task is to help the client to handle the uncertainty that they encounter. The 'you' of the analysis is the client, not the statistician. Our journals, and perhaps our practice, have been too divorced from the client's requirements. In this I have been as guilty as any. But at least the theoretician has developed methods. Your task is to put them to good use.

References

Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. Chichester: Horwood.

Bartholomew, D. J. (1988) Probability, statistics and theology (with discussion). *J. R. Statist. Soc.* A, **151**, 137–178.

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

Berger, J. O. and Wolpert, R. L. (1988) *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics.

Bernardo, J. M. (1999) Nested hypothesis testing: the Bayesian reference criterion. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Clarendon.

Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds) (1999) *Bayesian Statistics 6*. Oxford: Clarendon.

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

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

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973) Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. R. Statist. Soc.* B, **35**, 189–233.

DeGroot, M. H. (1970) *Optimal Statistical Decisions*. New York: McGraw-Hill.
Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–98.
Duckworth, F. (1998) The quantification of risk. *RSS News*, **26**, no. 2, 10–12.
Edwards, W. L., Lindman, H. and Savage, L. J. (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.
Eggleston, R. (1983) *Evidence, Proof and Probability*. London: Weidenfeld and Nicolson.
de Finetti, B. (1974) *Theory of Probability*, vol. 1. Chichester: Wiley.
———(1975) *Theory of Probability*, vol. 2. Chichester: Wiley.
Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
Healy, M. J. R. (1969) Rao's paradox concerning multivariate tests of significance. *Biometrics*, **25**, 411–413.
Jeffreys, H. (1961) *Theory of Probability*. Oxford: Clarendon.
Lehmann, E. L. (1983) *Theory of Point Estimation*. New York: Wiley.
———(1986) *Testing Statistical Hypotheses*. New York: Wiley.
O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
Onions, C. T. (ed.) (1956) *The Shorter English Dictionary*. Oxford: Clarendon.
Pearson, K. (1892) *The Grammar of Science*. London: Black.
Ramsey, F. P. (1926) Truth and probability. In *The Foundations of Mathematics and Other Logical Essays* (ed. R. B. Braithwaite), pp. 156–198. London: Kegan Paul.
Savage, L. J. (1954) *The Foundations of Statistics*. New York: Wiley.
———(1977) The shifting foundations of statistics. In *Logic, Laws and Life: Some Philosophical Complications* (ed. R. G. Colodny), pp. 3–18. Pittsburgh: Pittsburgh University Press.
Stein, C. (1956) Inadmissibility of the usual estimation of the mean of a multivariate normal distribution. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability* (eds J. Neyman and E. L. Scott), vol. 1, pp. 197–206. Berkeley: University of California Press.
Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Comments on the paper by Lindley

Peter Armitage (*Wallingford*)

Dennis Lindley has written so frequently, and so persuasively, about the principles of Bayesian statistics, that we scarcely expect to find new insights in yet another such paper. The present paper shows how wrong such a prior judgment would be. Lindley's concern is with the very nature of statistics, and his argument unfolds clearly, seamlessly and relentlessly. Those of us who cannot accompany him to the end of his journey must consider very carefully where we need to dismount; otherwise we shall find ourselves unwittingly at the bus terminus, without a return ticket.

I wrote 'those of us' because there must be many who, like me, sympathize with much of the Bayesian approach but are unwilling to discard a frequentist tradition which appears to have served them well. It is worth trying to enquire why this should be so. One possibility, of course, is that our reluctance is, at least in part, a manifestation of inertia, demonstrating a lack of courage or understanding. I must leave that for others to judge. I think, though, that there are sounder reasons for withholding full support for the Bayesian position. Lindley and I came to statistics during the 1940s, at a time when the subject was dominated by the Fisherian revolution. During the 19th century inverse probability had co-existed uneasily with frequentist methods by the use of flat priors, standard errors and normal approximations, results being interpretable by either mode of reasoning, albeit with occasional lack of clarity. Fisher had, it seemed, cleared the air by disposing of the need for inverse probability. Philosophical disputes, such as those with Neyman and E. S. Pearson, took place *within* the frequentist school, although Jeffreys and a few other pioneers maintained and developed the Laplace–Bayes framework. To many of us entering the field at that time it would have seemed bizarre to overturn such a powerful body of ideas. It is greatly to the credit of Lindley, and of a few of his contemporaries like Good and Savage, that they recognized the possibility that, as they might have put it, the Emperor had no clothes.

The great merit of the Fisherian revolution, apart from the sheer richness of the applicable methods, was the ability to summarize, and to draw conclusions from, experimental and observational data without reference to prior beliefs. An experimental scientist needs to report his or her findings, and to state a range of possible hypotheses with which these findings are consistent. The scientist will undoubtedly have prejudices and hunches, but the reporting of these should not be a primary aim of the investigation. Consider, for instance, one of the major achievements of medical statistics in the last half-century—the first study of Doll and Hill (1950) on smoking and lung cancer. They certainly had prior hunches, e.g. that air pollution was more likely than smoking to cause lung cancer, but it would have served no purpose to quantify these beliefs and to enter them into the calculations through Bayes's

theorem. There were indeed important uncertainties, about possible biases in the choice of controls and about the possible existence of confounding factors. But the way to deal with them was to consider each in turn, by scrupulous argument rather than by assigning probabilities to different models. That is an arbitrary example, but one that could be replicated by studies in a wide variety of applied fields.

This is not to deny the importance of prior beliefs in the weighing up of evidence, or especially in the planning of future studies, but rather to cast doubt on the need to quantify these beliefs with any degree of precision. As Curnow (1999) remarked, in connection with studies on passive smoking,

> 'Bayesian concepts must be fundamental to our way of thinking. However, quantifying and combining in any formal way the evidence on mechanisms with that from the epidemiological study are, in my view, impossible.'

To return, then, to Lindley's omnibus, I find that I should have dismounted by stage (b) in the list of stages (a)–(e) in Section 7. I believe that there are many instances of uncertainty which are best approached by discussion and further investigation and which do not lend themselves to measurement by probability. In that way I am absolved from the later need to dismount at stage (d). On further thought, though, I should perhaps have dismounted at (a), where, we are told, 'Statistics is the study of uncertainty'. This seems to claim too much for statistics, as indeed the author recognizes at the end of Section 2. I am surprised that he abandons the more traditional identification of statistics with the study of groups of numerical observations. Uncertainty still comes into the picture, by way of unexplained or 'random' variation, but this more modest view of our subject puts the emphasis on frequentist variation rather than the more ambitious Bayesian world view.

Frequentists are accustomed to receiving generous amounts of criticism from Bayesians about their incoherent practices. (Incidentally, I am not at all sure that a little incoherency is not a good thing. As Durbin (1987) implied, to look at a problem from irreconcilable points of view may generate useful insight.) Significance tests come in for especially hard knocks. Thus (Section 6), 'The interpretation of "significant at 5%" depends on n'. Well, it depends what you mean by 'interpretation'. In a rather obvious sense, the definition is independent of n. The possible results are arranged in some meaningful order and ranked by their probability on the null hypothesis. Irrespective of the sample size, a result that is significant at 5% comes beyond the appropriate percentile of the null distribution. What the objection means is that, on a particular formulation of prior probabilities for the non-null hypothesis, the Bayes factor comparing the two hypotheses will vary with n (Lindley, 1957). However, there is no reason why the non-null priors should be the same for different n. Large sample sizes are appropriate for the detection of small differences, and we might expect the non-null prior to be more concentrated towards the null for large rather than small n. With an adjustment of this sort the phenomenon disappears (Cox and Hinkley (1947), section 10.5).

Lindley sometimes seems to underestimate the ability of frequentist methods to cope with complexities. For instance (Section 6), 'meta-analysis is a difficult area for statistics'. As he says, this is merely the task of combining the results of many experiments. It has come rather late to some disciplines, such as clinical trials, because previously it was unusual to find many replicated studies that were sufficiently similar to be combined sensibly. But a frequentist analysis, combining estimates and not significance levels, is straightforward. In an earlier generation it was a standard exercise in agricultural trials (Yates and Cochran, 1938) and in bioassay (Finney (1978), chapter 14).

Again, he asserts (Section 14) that frequentist analysis is 'unable to cope' with the effects of changes in personal behaviour or Government policy on predictions of future numbers of cases of acquired immune deficiency syndrome. Yet the effects of changes in sexual practices and of improvements in therapy can be estimated by experiment or observation and built into the models that are used for such projections. It seems better to approach these problems by specific enquiries about each effect than to impose distributions of bias determined by subjective judgments.

Lindley's forceful presentation has led me to respond more robustly than I had expected. I respect the intellectual rigour of modern Bayesianism, and I acknowledge the influence that it has had in reminding statisticians that ancillary information is important, whether or not it is included in a formal analysis of the current data. In particular for decisions and verdicts a Bayesian approach seems essential, although here again a fully quantitative analysis may be unnecessary. However, I cannot agree that for the reporting of typical scientific studies a Bayesian analysis is mandatory. I have no objections to those who wish to follow that route, provided that they can make their conclusions comprehensible to their readers, but I wish to reserve the right to present my own conclusions in a different way. Unfortunately, such eclecticism is unlikely to find much support among the Bayesian community.

M. J. R. Healy (*Harpenden*)

In this paper Dennis Lindley sums up the experience of 50 years' advocacy of the subjective approach to statistical reasoning. It has been a long haul; he must at times have felt sympathy for Bishop Berkeley, of whose arguments it was said that they admitted no refutation but carried no conviction. Today though, thanks very largely to work by him and his students, Bayesian methodology is widely studied and the Royal Statistical Society's journals routinely carry papers in which Bayesian techniques are employed. Yet it remains true that the impact of this methodology on the vast amount of statistical analysis that is published in the scientific literature is essentially negligible. Almost every paper that I see as statistical adviser to a prominent medical journal contains the sentence 'Values of p less than 0.05 were regarded as significant' or its equivalent, and non-statistical referees regularly criticize submitted papers for an absence of power calculations or of adjustments of significance levels to allow for multiple comparisons. If the Fisher–Neyman–Pearson paradigm (Healy, 1999) is demonstrably unsatisfactory, as Professor Lindley claims to show, then large quantities of research data are being wrongly interpreted and a highly unsatisfactory situation exists. It seems to me that there may be more than one reason for this.

The first, and probably the least important, consists of weaknesses in the theoretical underpinning of Bayesian methods. One of these relates to the representation of ignorance, an area in which much work has been done. (It may be that we are never truly ignorant, but there are merits in enquiring what we should believe if we had been so.) Walley (1996) is particularly relevant here, and I must confess that I found the contributions to the discussion by Professor Lindley and some of his colleagues unconvincing.

Another issue stems from the fact that the bulk of statistical work is actually done by non-statisticians. This is as it should be; one of the responsibilities of the statistical profession (one that is not always lived up to) is that of making its insights available to research workers in all fields. Medical students today are exposed to statistical teaching in their preclinical years and later to text-books (not all written by statisticians) which lay down the standard approach of t, χ^2, r, Wilcoxon and the rest. If they wish to publish the results of research, they are liable to be encouraged, not to say compelled, to quote p-values and confidence limits—the paradigm that I have referred to is in full possession of the field. As statisticians we may come to follow Professor Lindley and to agree among ourselves that new and incompatible techniques must replace it, but how are we to explain this to our clients? Are we to apologize and suggest that they must unlearn all that we have been teaching them for many decades, that they must abandon their favourite computer packages? And, if we do, will they listen to us?

But the most severe problem, I suggest, is essentially a matter of psychology. Mankind in general longs for certainty, and the rise of natural science has been seen as a way of obtaining certain knowledge. We statisticians have pointed out that complete certainty is unobtainable, but we have maintained that the degree of uncertainty is quantifiable and objective (Schwartz, 1994)—we can be certain about how uncertain we should be. If we now insist on the personal subjective nature of uncertainty, how do we wish scientists to behave when they present their results? Are the conclusions to be preceded by the rubric 'In our opinion', with the implied parenthesis '(but you don't have to agree with us)'? We cannot even fall back on the objectivity of the data, since (as Professor Lindley has pointed out to me) the likelihood function itself depends on a model which is itself subjectively chosen. It may be that the inclusion of such a rubric would show a certain degree of realism—we can all remember papers to which our reaction was 'I don't believe a word of it'. But it must be admitted that it will not be welcomed by the scientific community as a whole, let alone by the general public.

Dennis Lindley's paper, like so many of his previous contributions, raises innumerable topics that are worthy of deep thought and discussion. There is no escaping the fact that statistics, unlike most disciplines, demands philosophical investigation. As practitioners we owe him a debt of gratitude for persuading us, unwilling as we may be, that such investigations must be pursued and for laying down one path that needs to be followed.

D. R. Cox (*Nuffield College, Oxford*)

It is 50 years since Dennis Lindley and I became colleagues at the Statistical Laboratory, Cambridge. Since then I have read most if not all of his work on the foundations of statistics always with admiration for its intellectual and verbal clarity and vigour. The present paper is no exception. It sets out with persuasiveness and aplomb the personalistic approach to uncertainty and individual decision-making. The ideas described are an important part of modern statistical thinking. A key issue, though, is whether they are the all-embracing basis of at least the more formal part of statistics or are to be taken as some part of an eclectic approach as I, and surely many others, have patiently argued; see, for example, Cox (1978, 1995, 1997) and Cox and Hinkley (1974).

On one point I believe that we are in total agreement. Bayesian in this context does not mean merely relying on a formal application of Bayes's theorem to produce inferences. Many of the applications involving flat priors or hyperpriors in a small number of dimensions can be regarded essentially as a combination of empirical Bayes methods and a technical device to produce sensible approximate confidence intervals implemented, for example, by Markov chain Monte Carlo methods. Provided that the relatively flat priors are not in a large number of dimensions, such investigations seem philosophically fairly neutral. Dennis Lindley's view is much more radical. It is the predominance of the constructive use of personalistic probability to synthesize various kinds of information, including that directly provided by data, into a comprehensive assessment of total uncertainty, preferably leading to a decision analysis. Flat priors have no role except occasionally as an approximation. The terminology 'Bayesian' is unfortunate, but I suppose that we are stuck with it.

Why is this unsatisfactory as the primary basis for our subject? In trying to discuss this in a brief contribution one is in the difficulty not merely of sounding more belligerent than is the intention but even more seriously of not having the last word! Also the paper is rich in specific detail on which comment is really desirable.

A major attraction of the personalistic view is that it aims to address uncertainty that is not directly based on statistical data, in the narrow sense of that term. Clearly much uncertainty is of this broader kind. Yet when we come to specific issues I believe that a snag in the theory emerges. To take an example that concerns me at the moment: what is the evidence that the signals from mobile telephones or transmission base stations are a major health hazard? Because such telephones are relatively new and the latency period for the development of, say, brain tumours is long the direct epidemiological evidence is slender; we rely largely on the interpretation of animal and cellular studies and to some extent on theoretical calculations about the energy levels that are needed to induce certain changes. What is the probability that conclusions drawn from such indirect studies have relevance for human health? Now I can elicit what my personal probability actually is at the moment, at least approximately. But that is not the issue. I want to know what my personal probability ought to be, partly because I want to behave sensibly and much more importantly because I am involved in the writing of a report which wants to be generally convincing. I come to the conclusion that my personal probability is of little interest to me and of no interest whatever to anyone else unless it is based on serious and so far as feasible explicit information. For example, how often have very broadly comparable laboratory studies been misleading as regards human health? How distant are the laboratory studies from a direct process affecting health? The issue is not to elicit how much weight I actually put on such considerations but how much I ought to put. Now of course in the personalistic approach having (good) information is better than having none but the point is that in my view the personalistic probability is virtually worthless for reasoned discussion unless it is based on information, often directly or indirectly of a broadly frequentist kind. The personalistic approach as usually presented is in danger of putting the cart before the horse. I hope that Dennis Lindley will comment on this. Is the issue in effect a nuance of interpretation or, as I tend to think, a point of principle?

Another way of saying this is that we can put broadly three requirements on a theory of the more formal parts of statistics: that it embraces as much as possible in a single approach, that it leads to internally consistent (coherent) consequences and that it meshes well with the real world (calibration). Now the personalistic approach scores extremely well on the first two points. My difficulty is that I put very large, indeed almost total, weight on the third. If there were to be a choice between working self-consistently and being in accord with the real world, and of course we would like to do both, then I prefer the latter. The frequency-based approach attempts, often rather crudely, to put that first.

Take a simple situation in Mendelian genetics in which some probabilities of $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ arise. Are they approximate representations of some biological phenomena that were going on long before anyone investigated them or are they to be interpreted as essentially the convergence of your personalistic probability in the face of a large amount of information? The second view is interesting but, to my mind, the former is the preferred interpretation and the essential reason why the probabilities are important. This is under a philosophical position that is close to naïve realism; there is a real world out there which it is our task to investigate and which shows certain regularities captured, in this case, by biological constants.

This leads to the conclusion that the elicitation of priors is generally useful mainly in situations where there is a large amount of information, possibly of a relatively informal kind, which it is required to use and which it is not practicable to analyse in detail. An informal summary by experts into a prior distribution may then be a fruitful approach. It carries the danger, however, that the experts may be

wrong and treating their opinion as equivalent to explicit empirical data has hazards. In any case settling issues by appeal to supposed authority, while sometimes unavoidable, is in principle bad. It is, of course, also possible that data are wrong, i.e. seriously defective, but this is open to direct investigation.

I understand Dennis Lindley's irritation at the cry 'where did the prior come from?'. I hope that it is clear that my objection is rather different: why should I be interested in someone else's prior and why should anyone else be interested in mine? There is a parallel question: where did the model for the data-generating process come from? This is no trivial matter, especially in subjects like economics. Here any sort of repetition is very hypothetical and, although some economists assert a solid theoretical knowledge base, this seems decidedly shaky. The reason for being interested in models is, however, clear. They are an imperfect but hopefully reasoned attempt to capture the essence of some aspect of the real physical, biological or social world and are in principle empirically at least partly testable. If we have a reasonably fruitful representation then in principle everyone is or should be interested in it.

The need for personal judgment, perhaps supremely in scientific research, is not in dispute. The formalization of this may be instructive in some situations. A central issue concerns the role of statistical methods (not the role of statisticians which is a different matter). I see that role as primarily the provision of a basis for mathematical representation of physical random phenomena and for public discourse about uncertainty.

The Bayesian formalism is an elegant representation of the evolution of uncertainty as increasingly more information arises. In its simplest form it is concerned with combining two sources of information. But one of the general principles in the combination of evidence is not to merge inconsistent 'data'. Consistency has usually to be interpreted in a probabilistic sense. Therefore we should face the possibility that the data and the other assessment (the prior) are inconsistent. (I am, of course, aware of the argument that no possibility should be excluded *a priori* but I cannot see that as a satisfactory way out.) Of course it may be that the data are flawed or being misinterpreted. But at least in principle something like a significance test seems unavoidable. I know that in principle we can reserve a small portion of prior probability to put on unexpected possibilities but surely we need also to represent what these are and this may be totally unknown. A complex set of data may show entirely unanticipated but important features. This is connected with the matter of temporal coherency on which comments would be welcome.

Are then p-values needed? It is interesting that for 50 years statisticians writing from a broadly frequentist perspective have criticized the overuse of significance tests (Yates, 1950). Indeed in some fields, notably epidemiology, conclusions are now primarily presented via approximate confidence limits which could be regarded as an approximate specification of a likelihood function, if that point of view were preferred. But in principle it seems essential to have a way of saying 'the data under immediate analysis are inconsistent with the representation suggested'. Now I agree with Dennis Lindley that it is necessary to have some idea of an alternative but not that it is necessary to formulate it probabilistically: desirable maybe, but necessary no. For example we may test for linearity without an explicit idea of the form of non-linearity that is appropriate. If the need arises we may then have to formulate specific new models but not otherwise.

The construction of overviews (so-called meta-analyses) via treating p-values as uncertainty measures is clearly a poor procedure if estimated effects and measures of precision are available on a comparable scale. But so also would be overviews in which measures of the degree of belief in some hypothesis were the only evidence available.

It seems to be a fundamental assumption of the personalistic theory that all probabilities are comparable. Moreover, so far as I understand it, we are not allowed to attach measures of precision to probabilities. They are as they are. A probability of $\frac{1}{2}$ elicited from unspecified and flimsy information is the same as a probability based on a massive high quality database. Those based on very little information are unstable under perturbations of the information set but that is all. This relates to the previous point and to the usefulness of such measures for communication.

Forecasting of acquired immune deficiency syndrome (AIDS) is mentioned as an exemplar of model uncertainty. Now the initial report on AIDS in the UK discussed sources of uncertainty and stated explicitly that model uncertainty was the major source. Indeed the message was rammed home by a front cover which showed several quite different forecasts as curves against time. Would it have helped to put probabilities on the different models and to have produced an overall assessment? I suppose that it is a matter of judgment but it seems to me that this would have been a confusing and misleading thing to do and would have hidden rather than clarified the issues involved. To put the point gently, the idea that only Bayesians are concerned about model uncertainty is wrong.

Dennis Lindley puts decision-making as a primary objective. Now I agree that such questions as why is this issue being studied and what are the consequences of such and such conclusions must always be considered whatever the field of study. At the same time I have rarely found quantitative decision analysis useful although I certainly accept this as a limitation of personal experience and imagination. For example, in the AIDS predictions mentioned above, the summary of the forecasts into a recommended planning basis was based on an informal decision analysis based on the qualitative idea that it was better to overpredict, leading to an overprovision of resources, than to underpredict, leading to a shortfall. It would have been difficult to put this quantitatively other than as a very artificial exercise via a series of sensitivity analyses. In most of the applications that I see, the role of statistical analysis is, in any case, to provide a base for informed public discussion.

Over the design of experiments, I do not see that as primarily the preserve of statisticians and it is important that most experiments are done to add to the pool of public knowledge of a subject and therefore should, for their interpretation at least, not be too strongly tied to the priors of the investigator. With a different interpretation of the word public this applies to industrial experiments also.

I do not understand the comment that theory makes a prediction about the proportion of hypotheses rejected at the 5% level in a significance test that are in fact false. How can theory possibly show anything of the sort? They may all be false or all (approximately) true depending on what we chose to investigate. I agree that it is the case that many assessments of uncertainty underestimate the error involved but this is for a variety of empirical reasons, the use of models ignoring certain components of variance or biases (which statisticians surely do not in general ignore), real instabilities in the effects under investigation and so on.

My attitude may be partly a reflection of a lack of mastery of current computational procedures but I am deeply sceptical of the advice of Savage to take models as complex as we can handle. This seems a recipe for overelaboration and for the abandonment of an important feature of good statistical analyses, namely transparency, the ability to see the pathways between the data and the conclusions.

I agree that prediction is underemphasized in many treatments of statistics and that the test of a representation of data is its ability to predict new observations or aspects of the original data not used in analysis. But this does not mean that prediction is necessarily the right final objective. We are not interested in estimating the velocity of light to predict the next measurement on it.

In conclusion, and not directly a comment on the paper, I want to object to the practice of labelling people as Bayesian or frequentist (or any other 'ist'). I want to be both and can see no reason for not being, although if pushed, as I have made clear, I regard the frequentist view as primary, for most if not virtually all the applications with which I happen to have been involved. I hope that by combining this view with a high regard for the present paper I am not committing an ultimate sin: incoherency.

J. Nelder (*Imperial College of Science, Technology and Medicine, London*)

Recently (Nelder, 1999) I have argued that statistics should be called statistical science, and that probability theory should be called statistical mathematics (not mathematical statistics). I think that Professor Lindley's paper should be called the philosophy of statistical mathematics, and within it there is little that I disagree with. However, my interest is in the philosophy of statistical science, which I regard as different. Statistical science is not just about the study of uncertainty, but rather deals with inferences about scientific theories from uncertain data. An important quality about theories is that they are essentially open ended; at any time someone may come along and produce a new theory outside the current set. This contrasts with probability, where to calculate a specific probability it is necessary to have a bounded universe of possibilities over which the probabilities are defined. When there is intrinsic open-endedness it is not enough to have a residual class of all the theories that I have not thought of yet. The best that we can do is to express relative likelihoods of different parameter values, without any implication that one of them is true. Although Lindley stresses that probabilities are conditional I do not think that this copes with the open-endedness problem.

I follow Fisher in distinguishing between inferences about specific events, such as that it will rain here tomorrow, and inferences about theories. For inferences about events, Lindley's analysis is persuasive; if I were a business man trying to reach a decision on whether to invest a million pounds in a project, I would act very much as he suggests. In analysing data relative to one or more scientific theories, I would wish to present what is objective and not to mix this with subjective probabilities which are derived from my priors. If the experimenter whom I am working with wishes to combine likelihoods with his own set of weights based on his (doubtless more extensive) knowledge then he is at liberty to do so; it is not my job to do it for him. However, if he wishes to communicate the results to other scientists, it would be

better, in my view, to stay with the objective part. (This paragraph is heavily dependent on ideas of George Barnard.)

General ideas like exchangeability and coherence are fine in themselves, but problems arise when we try to apply them to data from the real world. In particular when combining information from several data sets we can assume exchangeability, but the data themselves may strongly suggest that this assumption is not true. Similarly we can be coherent and wrong, because the world is not as assumed by Lindley. I find the procedures of scientific inference to be more complex than those defined in the paper. These latter fall into the class of 'wouldn't it be nice if', i.e. would it not be nice if the philosophy of statistical mathematics sufficed for scientific inference. I do not think that it does.

A. P. Dawid (*University College London*)

It is a real pleasure to comment on this paper. Dennis Lindley has been one of the most significant influences on my professional life, and his words are always worth reading carefully and taking to heart. It is in no way a criticism to say that I recognize, in the current work, ideas that Dennis has been promoting throughout the 30 years and more that I have known him—these things are still worth saying, perhaps now more than ever. For those who wish to read more of his penetrating and thought-provoking analyses, I particularly recommend Lindley (1971) and Lindley (1978), which contain some fascinating and educational examples of the differences between the frequentist and the Bayesian approaches to problems and clearly point up the logical difficulties that can arise when we do not conform to the principles of Bayesian coherence.

A case for Sherlock Bayes?

A recent and very important real example of this has arisen in the area of forensic identification, a problem area to which Lindley made some important early contributions (Lindley, 1977).

We are asked to compare two cases. The following details are common to each case. A murder has been committed, and a DNA profile, which can be assumed to be that of the murderer, has been obtained from blood at the scene of the crime. A suspect has been apprehended, and a DNA profile obtained from his blood. The two profiles match perfectly. The probability of this event, if the suspect is innocent, is some small number P—a realistic value might be $P = 10^{-6}$. (It is assumed that a match is certain if he is guilty. Then smaller values of the 'match probability' P may reasonably be taken as expressing stronger evidence against the suspect.) There is no other directly relevant evidence.

The difference between the two cases is that in case 1 the suspect was picked up at random, for completely unrelated reasons, and, on being tested, was found to match the DNA from the scene of the crime, whereas in case 2 a search was made through a computer database containing the DNA profiles of a large number N (perhaps $N = 10000$) of individuals, and the suspect (and no-one else in the database) was found to match.

The question to be addressed is 'In which of these two cases is the evidence against the suspect stronger?'. (Note that this question relates only to the strength of the evidence; we are not concerned with the possibility that the prior probabilities in the two cases might reasonably be different.)

The defence counsel argues, with mathematical correctness, that, because of the 'multiple testing' that has taken place in case 2, the probability of finding a (single) match in the database, if the true murderer is not included in it, is around NP (the probability of finding more than one match is entirely negligible). Since this match probability NP for case 2 is very substantially larger than the match probability P for case 1, that means that the evidence against the suspect in case 2 is *very much weaker*.

The prosecution counsel points out that, in case 2, one consequence of the search was to eliminate the other $N - 1$ individuals in the database as possible alternative suspects, thus *increasing* the strength of the evidence against the suspect—albeit by a typically negligible amount—above that for case 1.

Readers may like to assess whether they are intuitive Bayesians or intuitive frequentists by deciding which of these two arguments they prefer. Although both are based on probability arguments, only one of them is in accordance with the coherent use of probability to measure and manipulate uncertainty. Instead of identifying which this is, I shall just give a hint: consider the extreme case that the database contains records on everyone in the population.

For further reading on this problem, see Stockmarr (1999) and Donnelly and Friedman (1999); for more general application of coherent Bayesian reasoning to forensic identification problems, see Balding and Donnelly (1995) and Dawid and Mortera (1996, 1998).

Thomas Bayes in the 21st century

I share Lindley's view that, much as the tremendous recent expansion of interest in Bayesian statistics is to be welcomed and admired, its emphasis on computational aspects can sometimes stand in the way of a fuller understanding and appreciation of the Bayesian approach. It was the deep logical and philosophical conundra that beset the making of inductive inferences from data that attracted me into statistics in the first place and have exercised me ever since. But I have always been disappointed that so few other statisticians seem to share my view of statistics as 'applied philosophy of science', and even that small number seems to be dwindling fast. On the positive side, there are increasing numbers of researchers in artificial intelligence and machine learning who are taking foundational issues extremely seriously and are conducting some very original and important work. It is ironic that, as statisticians devote more of their effort to computing, so computer scientists are applying themselves to statistical logic.

When I was starting out, Bayesian computation of any complexity was essentially impossible. We could handle a few simple normal, binomial and Poisson models, and that was it. Whatever its philosophical credentials, a common and valid criticism of Bayesianism in those days was its sheer impracticability. Indeed, when I was engaged in organizing the first meeting on 'Practical Bayesian statistics' (sponsored by what was then still the Institute of Statisticians) in Cambridge in 1982, it was still possible for an eminent statistician to write to the Institute's newsletter suggesting that this was 'a contradiction in terms': an extreme and biased judgment, perhaps, but with a grain of truth. So, as we could not compute, we had to devote ourselves instead to foundational issues.

How things have changed! With the availability of fast computers and sophisticated computational techniques such as Markov chain Monte Carlo sampling, Bayesians can now construct and analyse realistic models of a degree of complexity which leaves most classical statisticians far behind. This power and versatility is itself a very strong argument for doing statistics the Bayesian way—far stronger, perhaps, than deep consideration of the logic of inference. But it would be sad if this practical success were at the expense of a clear understanding of what we are doing, and why we are doing it.

What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria. I have heard people who should know better use this 'dullness' as an argument against Bayesianism. One might as well complain that Newton's dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy's epicyclic system.

Nevertheless, the Ptolemaic temptation is difficult to resist and is apparent in much neo-Bayesian work, which struggles hard to escape from the restrictive confines of the fully coherent subjectivist Bayesian paradigm, dreaming up instead its own new and clever tricks. I regard this as a seriously wrong direction. All my experience teaches me that it is invariably more fruitful, and leads to deeper insights and better data analyses, to explore the consequences of being a 'thoroughly boring Bayesian'. Without a clear appreciation of what being coherent entails, and the guidance that a strict Bayesian framework supplies, it is all too easy to fall into erroneous and misguided ways of formulating problems and analysing data.

J. F. C. Kingman (*University of Bristol*)

This paper is of great importance. If 'philosophy' is read as 'general principles', the author is laying down the general principle that the output from any statistical analysis should consist of a number of probability statements. These are subjective in the sense that they depend on assumptions made by the analyst and stated in the report, and another analyst with different prejudices will produce different conclusions.

I use the word 'analyst' rather than 'statistician' because the argument, if it is valid at all, may apply not just to statistical method but to any reported research in which uncertainty plays a part. Thus Professor Lindley is calling for a revolution in the way that research in general is carried out and reported, and is doing so on the basis of very simple arguments of coherence. If we do not follow his advice, he can make money systematically from us by asking us to bet on our conclusions.

I first encountered the clarity and deceptive simplicity of Professor Lindley's exposition as a Cambridge freshman listening enthralled to his introductory course on statistics. Much of what he taught us then he would now recant, but the way in which the complexities of an uncertain world were fitted into an elegant and convincing theory was deeply impressive. Perhaps mathematicians select themselves

by this desire to reduce chaos to order and only learn by experience that the real world takes its revenge.

The most common reason for scepticism about the Bayesian approach is the apparently arbitrary nature of the prior distribution $p(\theta)$, but I worry even more about the 'model' $p(x|\theta)$, which so many statisticians, Bayesian or otherwise, seem to take for granted. Just what evidence do we need to convince us that a particular model, with a particular meaning for the parameter θ, is or is not appropriate to a particular problem?

Special aspects of this question have of course been studied in theoretical terms, but in practice many statisticians make a conventional choice, often based on mathematical or computational convenience. This habit seems to me to be based on a feeling that, although statistical inference is difficult and controversial, the probability calculus at least is a firm foundation that need not be questioned.

Mathematicians since Kolmogorov have connived by presenting the mathematics of probability as following irresistibly from the general theory of measure and integration, but the internal consistency of the mathematics is no guarantee that it applies to any real situation. Philosophers warn of the dangers of attaching firm meaning to *any* probability statement about the world, and the fact that such statements are undeniably useful to (for instance) the designers of telephone systems should not lead us to an uncritical reliance on what is in the end only a collection of mathematical tautologies.

One example must suffice. We teach our students that two events are (statistically) independent when the probability that they both occur is the product of their probabilities. We then forget the adverb, and assume that, if we cannot see any causal link between two events, the multiplication *law* must apply. At the level of constructing a plausible model, this is a reasonable procedure, so long as the model is then tested. But how do we test the sort of assertion that is made about the safety of nuclear power-stations, that the probability of disaster is 10^{-N}, where N is some very large number. The assertion is based on many applications of the 'multiplication law', ignoring the fact that the justification of the law is inherently circular.

So probability statements are dangerous currency even before we try to infer them from dirty data. We must distrust the prophets who can sum up all the complexities in a few simple formulae. But such scepticism does not absolve statisticians from asking what the general principles of their subject are. If we do not accept Professor Lindley's prescription, what alternative do we have?

David J. Bartholomew (*Stoke Ash*)

It is a pleasure to comment on this lucid and authoritative exposition of subjective Bayesianism. There is much in the paper with which I whole-heartedly agree, but I shall focus on the points of difference. I agree that uncertainty and variability are at the heart of statistics. Unlike the author, however, I regard variability as the more fundamental. Data analysis then comes first and does not have to be justified later as a tool for model selection. Uncertainty arises naturally, but secondarily, when we need to think about $p(y|x)$ or $p(\theta|x)$.

My main point concerns modelling. Debates on inference have often treated the model as given and so focused on the prior distribution. How the model is chosen is much more important for the philosophical foundations of statistics. Lindley argues for the largest possible model. What happens when we push this to the limit and try to imagine a truly global model for the whole cosmos? In the beginning the model's $\mathbf{x}$ would have to include literally everything that could be observed. At that stage K, the background knowledge, is an empty set. How could we then assign a prior without any background knowledge? If this is impossible, how does the journey to knowledge ever start?

But if we allow that, somehow or other, it did start what matters now is how we proceed; just how large does the 'world' of the model have to be? How do we cope with the fact that different models may have the same observational consequences? Why should two scientists ever agree, no matter how much data they have in common, if they are operating with different, but equally well-supported, models? In any case, any realistic world model is underdetermined and so certainty is beyond our reach.

Experimental science attempts to solve the 'small' world problem by controlling all extraneous variables. R. A. Fisher took that idea further by using randomization to ensure that the extraneous effects were controlled on average. It is a weakness of the subjective Bayesian philosophy that it has no place for randomization and thus no way of making valid unconditional inferences in a small world. Instead it is left to flounder on the slippery slope of what looks suspiciously like an infinite regress.

Finally, it is the personal focus of the philosophy which makes me most uneasy. The pursuit of knowledge—inference and decision-making—is a collective as well as an individual activity. It is not, essentially, about what it is rational for you or me to believe and do, but about what claim there is on us all, collectively, to believe something or to act in a particular way. Lindley recognizes this distinction in

decision-making but it is equally relevant in inference. Inference is conditional on the model and without agreement on where the journey starts there is no guarantee that we shall all arrive at the same destination.

Attractive though it is, the author's world of discourse seems too small and self-contained to be the last word.

A. O'Hagan (*University of Sheffield*)

I congratulate Dennis Lindley for his elegantly written paper, that so lucidly covers an enormous range of fundamental topics. I particularly liked the section on 'data analysis' in Section 10. The idea, that we should condition on whatever summaries $t(x)$ of the data have been used in building or checking the model, is a real insight. It clearly covers the case where we use part of the data as a 'training sample', reserving the remaining data for confirmatory analysis, and so links to the use of partial Bayes factors (O'Hagan, 1995). It will, of course, be more difficult to apply following more loosely structured 'data analysis'.

I also applaud the emphasis, in the final section, on the need for research into methods of assessing (or eliciting) probability distributions.

Lindley misses an opportunity, however, to show how the Bayesian approach clarifies the concept of a nuisance parameter. He says, at several points, that it may be 'necessary' to introduce nuisance parameters α in addition to the parameters of interest θ. In what sense is this 'necessary'?

Lindley introduces, in Section 9, the example of a doctor needing to give a prognosis y for a patient based on observations x from previous patients. He says that this could be done simply by assessing the predictive distribution $p(y|x)$, but that this is 'usually difficult'. He then asserts that 'a better way to proceed ... is to study the connections between x and y, and the mechanisms that operate'. This argument only works if we recognize the limitations of practical probability assessment. To assess $p(y|x)$ directly is not just 'difficult' but likely to be very inaccurate, whereas constructing it indirectly via other assessments (a model) and the laws of probability is both more accurate and more defensible to others. In O'Hagan (1988) I develop this idea of 'elaboration' as a fundamental tool of probability measurement.

Taking this view, nuisance parameters are desirable (if not absolutely 'necessary') to achieve sufficiently accurate and defensible assessments; we can be confident of assessing $p(y|x, \alpha, \theta)$ but not of assessing $p(y|x, \theta)$.

Finally, I recognize that in such a concise survey as this it is necessary to make some judicious simplifications, but there are a few places where Lindley risks damaging his argument by being simplistic rather than simplified.

(a) The example in Section 8 of non-conglomerability when ignorance is represented by an improper uniform distribution is not a good one, because the conditional probabilities are not defined. The conditioning events have zero probability.
(b) At the end of Section 8, it is not true that consensus will be reached if, as stated in the next sentence, we admit subjectivity about the likelihood.
(c) At the end of Section 11, the possibility of tactical voting is overlooked. One's utility is not simply a matter of the number of votes cast for the 'best' candidate.

David J. Hand (*Imperial College of Science, Technology and Medicine, London*)

I would like to congratulate Professor Lindley on a masterfully clear exposition of the Bayesian perspective. The arguments that he presents for adopting this approach to inference are difficult to refute. Nevertheless, I must take issue with the paper, and my disagreement begins with the title. What is described in the paper is a philosophy of statistical inference, not a philosophy of statistics. As such, it ignores much which should properly be regarded as within the orbit of statistics.

Lindley suggests, in Section 2, that statistics is the study of uncertainty. This is certainly one of the most important aspects of statistics, perhaps the largest part, but it does not define it. At the very least, it leaves out description, summarization and simplification when the data are not uncertain and the aim is not inference, as arises, for example, when the complete population is available for analysis. Would Professor Lindley claim that data analytic tools such as multidimensional scaling and biplots are not statistical tools? Would he claim that the clustering of chemical molecules, when data are available for the entire population of a given family of molecules, is not statistics? Would he claim that clustering microarray gene expression data is not statistics?

This would not be a serious issue if it were merely a matter of terminology. But it is not. It goes

further than this and has implications for how the discipline of statistics as a whole is perceived. In my view, the narrow view of statistics which it implies has contributed to the fact that other data analytic disciplines have grown up and adopted subject-matter, *kudos* and resources which are more appropriately regarded as belonging to statistics. Again, this would not matter if it were only a question of hurt pride. But, again, there is more involved. In particular, it means that the elegant tools for handling uncertainty which have been developed by statisticians have not always been adopted by others concerned with data analysis, and have therefore not been applied to problems which could benefit from them. For example, database technologists have not always appreciated the need for inferential methods (a case in point being the analysis of supermarket transaction data, where the discovery of a relationship in the data has been taken at face value, with no explicitly articulated notion of an underlying population from which the data were drawn). A second example is the case of fuzzy logic. The underlying logic here has not been uniquely defined, and the area certainly lacks the elegant rigour of the Bayesian inferential strategy described by Professor Lindley. But the methods now attract a huge following. This following tends to come from the computational disciplines—where, because of the narrow view of statistics described above, statistical inferential methods have not been adopted as fundamental. A third example is computational learning theory, which began by assuming that the classes in supervised classification problems were, in principle at least, perfectly separable and has only recently begun to struggle with the more realistic non-separable case which statisticians take for granted.

Sometimes there is a convergence—artificial neural nets are a most important recent example, and recursive partitioning tree methods, developed in parallel by the statistical and machine learning communities, are another. When this happens significant synergy can result from the integration of the different perspectives. It is a pity, and detrimental to the rate of scientific progress, that a period of separation has to exist at all.

One issue on which I would welcome Professor Lindley's comments is the issue of what I call 'problem uncertainty'. The inferential strategy outlined in the paper captures model uncertainty and sampling uncertainty, but real problems often have an extra layer of uncertainty, in that the question that the researcher is trying to answer is not precisely defined. An obvious illustration lies in the need to operationalize measurements: in physics we may have a good idea that our measuring instruments match our conceptual definition of a variable, but in many other domains things are not so clear cut. Our model may predict a good outcome if we measure a response in one way, but what if there is a disagreement about the best way to measure the response? An extreme example would be quality-of-life measurement. Similarly, in a clinical trial, the response to a treatment may be measured in different ways. And in classification problems, for example, it is not always clear how to weight the relative costs of the different kinds of misclassification. How should we take into account this kind of uncertainty?

On a minor point, if I disagree with Professor Lindley about the scope of statistics, I perhaps disagree with him even more about the scope of literature. Analyses of word counts may 'help to identify the author of an anonymous piece of literature' (Section 18), but they do not say anything about literature *per se*.

I would like to end on a note of agreement. Lindley remarks, in his final paragraph, that 'Our journals ... have been too divorced from the client's requirements'. This seems too painfully to be the case. The focus seems to be increasingly on narrow technical advance into increasingly specialized areas, with greater merit being awarded to work which is more abstract and more divorced from the realities of data. Statistics has enough of an image problem to overcome, without our gratuitously aggravating it. I am painfully reminded of Ronald Reagan's remark, that 'Economists are people who see something work in practice and wonder if it would work in theory'. I would hate statisticians to be tarred with the same brush.

George Barnard (*Colchester*)

Space does not permit listing the many points of disagreement, and some points of agreement, between me and my friend Professor Lindley. My central objection to probability as the *sole* measure of uncertainty is the rule $\Pr(H) + \Pr(\text{not } H) = 1$. If H is a statistical hypothesis that is relevant to a given data set E it must specify the probability $\Pr(E|H)$ of E. But the mere assertion that H is false leaves $\Pr(E|H)$ wholly unspecified. It is only when given a particular *model*, i.e. a specified collection M of hypotheses, that we are entitled to equate 'not H' with 'some other hypothesis in M'. Our model may be wrong, and the primary function of traditional p-values is to point to this possibility. If M is wrong, in repeated experimentation p will shrink to 0.

Given that M is accepted, our statistical problem becomes that of weighting the evidence for any one

H in M against that for any other H' in M. This is done by calculating the likelihood ratio

$$W = L(H \text{ versus } H'|E) = \Pr(E|H)/\Pr(E|H').$$

In any long series of judgments between pairs H and H', if we choose H rather than H' when L exceeds w and choose H' rather than H when L is less than $1/w$, leaving our choice undetermined when L falls between $1/w$ and w, correct choices will outnumber incorrect choices by at least w:1. For important choices we might fix $w = 100$, insisting on more data when w falls between 100 and $1/100$. For less important choices we might be willing to take $w = 20$. The more important our choice, the more data we may need to collect.

Likelihoods cannot always be added. But if with data E giving $L(\alpha, \beta|E)$ we are interested in α but not in β then, provided that the data themselves are reasonably informative about β, adding $L(\alpha, \beta)$ over β values is permissible as a reasonable approximation. Nowadays desk-top computers allow us to overview $L(\alpha, \beta)$ and it is easy to *see* whether, *for the data to hand*, such an approximation is permissible.

Fisher never had a desk-top computer. But in every edition of *Statistical Methods for Research Workers* he said that *likelihood was the measure* of credibility for inferences. To keep to statistical methods that were actually usable in his day he had to overstress p-values. I am sure that Fisher would change his mind today, and I hope that Professor Lindley may be persuaded to do the same.

Brad Efron (*Stanford University*)

The likelihood principle seems to be one of those ideas that is rigorously verifiable and yet wrong. My difficulty is that the principle rules out many of our most useful data analytic tools without providing workable substitutes. Here is a bootstrap story, not entirely apocryphal, to illustrate the point.

A medical researcher investigating a new type of abdominal surgery collected the following data on the post-operative hospital stay, in days, for 23 patients:

1 2 3 3 3 3 4 4 4 4 5 5
5 5 6 6 7 7 8 9 10 16 29.

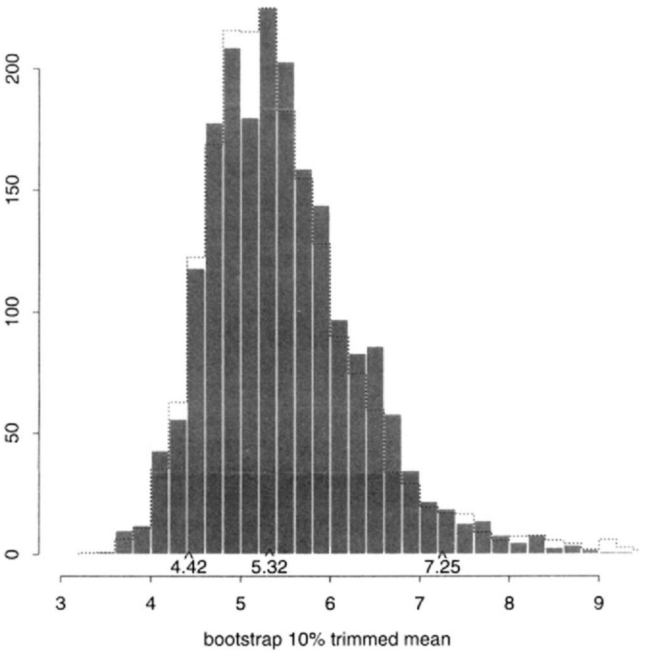

Fig. 1. 2000 bootstrap replications of the 10% trimmed mean for the hospital stay data: ·····, Bayesian bootstrap

Following a referee's advice she had summarized the data with its 10% trimmed mean, 5.35, but wanted some formula for the estimate's accuracy.

To help to answer her question I drew 2000 independent bootstrap samples, each comprising 23 draws with replacement from the data above, and for each sample computed the 10% trimmed mean. The histogram of the 2000 bootstrap trimmed means is shown in Fig. 1. From it I calculated a bootstrap standard error of 0.87 and a nonparametric 90% bootstrap approximate confidence interval [4.42, 7.25]. It was interesting to notice that the interval extended twice as far above as below the point estimate, reflecting the long right-hand tail of the bootstrap histogram.

This is exactly the kind of calculation that is ruled out by the likelihood principle; it relies on hypothetical data sets different from the data that are actually observed and does so in a particularly flagrant way. Some of the bootstrap samples put more than 10% sample weight on the two largest observations, 16 and 29, giving them influence on the 10% trimmed mean that they do not exert in the original data set. As a matter of fact this effect accounts for most of the long right-hand tail in the histogram.

It is not as though Bayesian theory does not have anything to say about this example. We might put an uninformative Dirichlet prior on the class of probability distributions entirely supported on the 23 observed values, as suggested in Rubin (1981) and in chapter 10 of Efron (1982), and calculate the posterior distribution of the population 10% trimmed mean. Interestingly, this posterior distribution agrees closely with the bootstrap histogram—I have indicated it by the dotted histogram in Fig. 1.

But of course this is not a genuine Bayesian analysis; it is empirical Bayesian, using the data to guide the formation of the 'prior' distribution. A well-known quote of Professor Lindley says that nothing is less Bayesian than empirical Bayes analysis. Does he still feel this way? My feeling is that good frequentist procedures are often carrying out something like an 'objective' Bayesian analysis, as suggested in Efron (1993), and that maybe this hints at a useful connection between the realities of data analysis and the philosophic cogency of the Bayesian argument.

D. A. Sprott (*Centro de Investigación en Matemáticas, Guanajuato*)

Some of my doubts about the application of the ideas in this paper to scientific inference are listed briefly below.

(a) This paper relegates statistical and scientific inference to a branch (probability) of pure mathematics, where inferences are deductive statements of implication: if H_1 then H_2. This can say nothing about whether there is reproducible objective empirical evidence for H_1 or H_2, as is required by a scientific inference. Scientific inference transcends pure mathematics.

(b) In particular, Bayes's theorem (1) requires that *all* possibilities $H_1, H_2, \ldots, H_k$ be specified in advance, along with their prior probabilities. Any new, hitherto unthought of hypothesis or concept H will necessarily have zero prior probability. From Bayes's theorem, H will then always have zero posterior probability no matter how strong the empirical evidence in favour of H.

(c) This demonstrates the necessity of presenting the likelihood function to summarize the objective experimental evidence. But what if the likelihood function flatly contradicts the prior distribution, leading to a posterior distribution that flatly contradicts both the prior distribution and the likelihood function? Surely contradictory, conflicting, items cannot merely be routinely combined.

(d) Likelihoods are point functions whereas probabilities are set functions. Likelihood therefore measures the relative plausibility of two specific values, $\theta':\theta''$, of a continuous parameter θ. The probability, however, of each specific value, θ' and θ'', is 0. Probability measures the uncertainty of intervals. The practical value of using likelihood supplemented by probability (if possible) to measure uncertainty is illustrated by Díaz-Francés and Sprott (2000).

(e) I do not see how the injection of subjective beliefs into experimental evidence (Section 8) can be justified. Beliefs are necessary in designing experiments. To inject them into the analysis of the objective data could lead to proof by assumption or belief, or to the combination of contradictory items as in (c) above. The beliefs may just be plainly wrong and should be rejected, not combined, however 'incoherent' this would be. In any case the likelihood should be presented separately as a summary of the empirical evidence, uncontaminated by beliefs. As put by Bernard (1957), page 23,

'I consider it, therefore, an absolute principle that experiments must always be devised in view of a

preconceived idea .... As for noting the results of an experiment, ... I posit it similarly as a principle that we must here, as always, observe without a preconceived idea.'

Ian W. Evett (*Forensic Science Service, London*)

I am grateful for this opportunity to comment on the paper by Professor Lindley. He and I have known each other for over 25 years and it would be difficult for me to exaggerate the effect that he has had on my thinking.

His view, like that of the majority of statisticians, is that of the mathematician. I am, first and foremost, a scientist—indeed, a forensic scientist. My perspective is quite different and might be considered iconoclastic, renegade even, to readers of this journal.

It is appropriate that Dennis should mention that there appears not to be a strong association between statistics and physics. I took my first degree in physics and my introduction to statistics came in my first year. It was a two-hour lecture on 'errors of observation and their treatment'; the lecturer was so proud of it that he recorded himself for future use. I understood none of it. My second year included a course on statistics from a real live academic statistician. I am not exaggerating when I say that I found it completely mystifying. Now clearly, I had to do *something* in my practical experiments to indicate the extent of the uncertainties in any estimates or other inferences that I drew from my observations. But that was not really a problem, because it soon became clear that my supervisors, to say nothing of my fellow students, understood no more of the finer points of statistics than I did! A good amount of fudging to round off one's experimental reports was quite enough to satisfy the most scrupulous demonstrator.

Later—by that time a practising forensic scientist—I was sufficiently fortunate to study statistics full time in a post-graduate course at Cardiff. Since then, I have spent most of my time working with forensic scientists of all disciplines on matters of inference and statistics. A large proportion of my time is devoted to training—with the emphasis on new entrants to the Forensic Science Service. And what do I find? Most graduate scientists have learned little of statistics other than a dislike of the subject. Even more importantly—they have no understanding of *probability*. A forensic scientist spends his or her time dealing with uncertainty. In courts, terms such as 'probable', 'unlikely' and 'random' are everyday currency. I have thought for a long time that if a forensic scientist is indeed a genuine scientist then he or she should understand probability.

Yet it is my impression that many statisticians are rather uncomfortable with the notion of probability. Certainly, there is plenty of talk about long run frequencies—but what about the probabilities of real world problems? Is it not a fundamental weakness that most texts and teaching schemes present *conditional* probabilities as something special? Even such a delightful text as *Chance Rules* (Everitt, 1999) talks quite happily about 'probability' before introducing 'conditional probability', almost as a new concept, in chapter 7. All probabilities are conditional. There is no such thing as an 'unconditional probability'. A probability is without meaning unless we specify the information and assumptions that it is based on. Whereas these assertions are to me obvious, I sense that many statisticians would dispute them.

Much progress has been made towards establishing the principles of forensic science through the Bayesian paradigm. There is little disagreement among Bayesians and frequentist alike that Bayes's theorem provides a logical model for the processing of evidence in courts of law. I have heard classical statisticians who have ventured into the field say things like 'I have nothing at all against Bayesian methods—indeed, I use them myself when they are appropriate'. But this is the cry of the toolkit statistician—a statistician who lacks a philosophy. In the world of science here lies the distinction between the scientist and the technician.

The new technology of DNA profiling has caught the headlines and brought many statisticians into the field. Much of this has been highly beneficial to the pursuit but there have been several eccentricities, arising from the classical view, that have confused rather than illuminated. An example of misguided statistical thinking is the notion of significance testing for Hardy–Weinberg equilibrium (HWE). We all know that the conditions for HWE cannot exist in the real world but what do we do when a new locus is implemented for DNA profiling? We set out to test the null hypothesis that HWE is true—even though we know that it is patently false! Why do we play these silly games? They do not help science and they do not help the advancement of statistics as a scientific discipline.

My view of statistics, as a scientist, is that there is the Bayesian paradigm and there is everything else—a hotchpotch of significance testing and confidence intervals that is at best peripheral to the scientific method. Yet this is what is taught to science undergraduates and most, like me, are mystified by it. In his concluding paragraph, Dennis says that statisticians 'have been too divorced from the client's

requirements'. Here is my request as a client: in the future, I require that all new science graduates should understand probability. And I am not talking about coin tossing and long run frequencies: I am talking about probability as the foundation of logical scientific inference.

Author's response

As explained at the end of Section 1, this paper began as a reprimand to my fellow Bayesians for not being sufficiently Bayesian, but it ended up by being a statement of my understanding of the Bayesian view. To many, this view acts like a red rag to a bull and I am most appreciative of the fact that the discussants have not been bullish but have brought forward reasoned and sensible arguments that carry weight and deserve respect. Limitations of space prevent every point from being discussed and omission does not imply lack of interest or dismissal as unimportant.

Many of the discussants are reluctant to abandon frequentist ideas and I agree with Armitage that this is not just inertia, though the difficulty all of us experience in admitting we were wrong must play a part. There are at least two solid reasons for this: frequency techniques have enjoyed many successes and, through the concept of exchangeability, share many ideas with the Bayesian paradigm. Against this there is the consideration that almost every frequentist technique has been shown to be flawed, the flaws arising because of the lack of a coherent underpinning that can only come through probability, not as frequency, but as belief. A second consideration is that, unlike the frequency paradigm with its extensive collection of specialized methods, the coherent view provides a constructive method of formulating and solving any and every uncertainty problem of yours. 30 years ago I thought that the statistical community would appreciate the flaws, but I was wrong. My hope now is that the constructive flowering of the coherent approach will convince, if not statisticians, scientists, who increasingly show awareness of the power of methods based on probability, e.g. in animal breeding (Gianola, 2000).

The paper, stripped to its bare essentials, amounts to saying that probability is the key to all situations involving uncertainty. What it does not tell us is how the probability is to be assessed. I have been surprised that, although the rules are so simple, their implementation can often be so difficult. Dawid provides a striking example that caused me much anguish when it was first encountered in the 1970s. Kingman's example of independence is sound and pertinent. I find it illuminating, when seeing yet another introductory text on statistics, to see whether the definition of independence is correct; often it is not. Some elementary texts do not even mention the notion explicitly and conditional probability is rarely mentioned. No wonder many of these texts are so poor.

In his perceptive comments, Cox may not have appreciated my view of the relationship between a statistician and their client. It is not the statistician who has the probabilities but the client; the statistician's task is to articulate the client's uncertainties and utilities in terms of the probability calculus, these being 'based on serious and, so far as feasible, explicit information' that the client has. This information may be based directly on data but often it uses deep understanding of physical and other mechanisms that are unfamiliar to the statistician. The idea of a statistician starting from a position almost of ignorance about mobile telephones and updating by Bayes's theorem, using what the expert says, is not how I perceive the process. The client has informed views and it is these that need to be quantified and, if necessary, modified to be made coherent. In the mobile telephones example there are presumably many clients including the manufacturers and environmentalists with opposing concerns. Although ideas exhibited in the paper do not directly apply to groups in opposition, and I do not know of any method that does in generality, they can assist, especially in exhibiting strange utility functions. It is interesting that, having expressed doubts about probabilities in the mobile telephone study, Cox concludes that

> 'the elicitation of priors is generally useful mainly in situations where there is a large amount of information, possibly of a relatively informal kind, which it is required to use and which it is not practicable to analyse in detail'.

Is not this a fair description of the study? Incidentally, the statistician's fondness for frequency data should not blind them to information to be had from a scientific appreciation of the underlying physical mechanism.

I agree with Cox that personalistic probability should be based on information; that is why it is always conditional on that information. But I do not see how he can claim that 'confidence limits ... could be regarded as an approximate specification of a likelihood function'. Observing r successes in n trials, a

likelihood function can be found but not a confidence limit because the sample space is undefined. Cox suggests that we can test a hypothesis without an alternative in mind. Yes, but whether the test will be any good depends on the alternative.

My emphasis, supported by Evett, on the conditional nature of probability, that it depends on two arguments, not one, has not been fully appreciated. For example, if hypotheses, H_1, H_2, ..., H_n are contemplated, with H their union, then all probabilities will be conditional on H, so the addition of H_{n+1} will only necessitate a change in the conditions to the union of H and H_{n+1}. This meets Barnard's objection about not-H, Nelder's point about likelihood and Sprott's point (b). Incidentally, it is interesting, though not unexpected that, apart from Barnard, no one attempts to demolish the arguments of Sections 1–6 leading to what has been called 'the inevitability of probability' and it is only when they mount the bus, in Armitage's happy analogy, that doubts enter. The proof of the pudding may be in the eating but the recipe counts also.

The identification of statistics with uncertainty has worried many, even though I explained that it was 'not in the sense of a precise definition' so Hand is allowed his exemptions, though even there it is well to recognize that, even with a complete population, a summary introduces uncertainty and the quality of a summary is judged by how little uncertainty it leaves behind. To answer Armitage, the reasons for choosing uncertainty as primary, rather than variability in the data, are first that it is the uncertainty of the parameter (or the defendant's guilt) that is primary and the data are aids to assessing it, and second that data need something more before they will be of value. In amplification of the second point, it is possible to have two $2 \times 2 \times 2$ tables, each with a control factor, another of effect and the last a confounding factor, both with the same numbers, in which the conclusions about the effect of the control are completely opposed (Lindley and Novick, 1981). Statistical packages that analyse data without context are unsound.

The suggestion of an eclectic approach to statistics, incorporating the best of various approaches, has been made. I would, with Evett, call it unprincipled. Why do adherents of the likelihood approach, part of this eclecticism, continue with their upwards of 12 varieties of likelihood, all designed in an attempt to overcome the failure of likelihood to be additive, a requirement easily seen to be essential to any measure of uncertainty? There is only one principle: probability. Why use a pejorative term like sin to describe incoherence? These eclectic people do not like principles, as is evident by their failure to consider them, instead concentrating on their perceptions of what happens when they are applied, often falsely. Efron worries about the likelihood principle, which is not surprising when the bootstrap has no likelihood about which to have a principle. The Bayesian view embraces the whole world, which is overambitious, and has to be reduced to small worlds, whereas the frequentist view restricts attention to a population. The bootstrap goes to the extreme and operates within the sample, eschewing reference to outside aspects and using *ad hoc* methods, like trimmed means, discussed in Section 12 within a coherent framework. Readers who are not already familiar with it might like to read the balanced discussion of the bootstrap in Young (1994) and, in particular, Schervish's remark that 'we should think about the problem'.

O'Hagan may be wrong when he says, in his point (b), that a consensus will not be reached if the likelihood is subjective. With two hypotheses and exchangeable data, your log-odds change, on receipt of data, by the addition of the sum of your log-likelihood ratios. Provided that the expectation of the ratios is positive under one hypothesis and negative under the other, the correct conclusion will be reached and hence consensus.

Bartholomew worries about this consensus in a wider context. I do not know how we get started; perhaps it is all wired in as Chomsky suggests grammar is. Interesting as this point is, it does not matter in practice because, when we are faced with quantities in a problem, they make some sense to us and therefore we have some knowledge of them. I may be unduly optimistic but I feel that if two people are each separately coherent, a big assumption, then a coherent appreciation of theory and experience will ultimately lead to agreement. It happens in science, though not in politics or religion, but are they coherent? A small correction to Bartholomew: Bayes does recognize what I have called a haphazard design (Lindley, 1982), and a convenient way to produce one is by randomization (after checking that the random Latin square is not Knut-Vik).

Hand raises the important question for our Society of why

> 'the elegant tools for handling uncertainty which have been developed by statisticians have not always been adopted'.

One reason may be that some statisticians, myself included, have not immersed themselves sufficiently

in the circumstances surrounding the data. Efron provides an example when, apart from telling us that the numbers refer to hospital stays after surgery, he just treats the 23 numbers as elements in the calculation: they might equally have referred to geophysics for all the bootstrap cares. We need to show more respect than we do for our clients. When I suggested this at a meeting recently, some members of the audience laughed and mentioned cranks who believed in alternative medicine as people who do not deserve respect. Ought we not to try to help all who come to us to express their ideas in terms of probability, to help them to become coherent and to respond more sensibly to data? Another reason for suspicion of statistics is that some of our methods sound, and are, absurd. How many practitioners understand a confidence interval as coverage of a fixed value, rather than as a statement about a parameter?

Cox defends the statistical analysis of acquired immune deficiency syndrome (AIDS) progression. My point here, perhaps not clearly expressed, for which I apologize, is that a frequentist approach can only lead to standard errors for estimates that use the frequentist variation that is present in the data. It cannot incorporate into its prediction other types of uncertainty. For example, it may happen that, impressed by media emphasis on AIDS, the public may act more cautiously in their sexual activities and, as a result, the incidence would decrease. This sort of judgment is outside the frequency canon and, although competent, dedicated frequentists will search for ways around the limitation, the Bayesian approach naturally incorporated both forms of uncertainty. In this connection, what is the objection to attaching probabilities to frequentist models to provide an overall Bayesian model? Hoeting *et al.* (1999) provides a good account. I find Cox's notion of prediction too narrow. Scientists want to estimate the velocity of light, not to predict a future measurement of velocity but to predict realities that depend on it, e.g. the time taken for the light from a star to reach us. In $p(y|x)$, y is not necessarily a repetition of x, but merely related to x; the relationship often being made explicit, as O'Hagan says, through a parameter. Of course, frequentists do not like prediction because it is so difficult within their paradigm, involving contortions akin to those with confidence intervals.

Several discussants raise the important issue of inconsistency, e.g. between a prior, to use the unfortunate term, and the likelihood. In this case, data have arisen which were unexpected. There are several possibilities. One is that an inspection of the data or discussion with a colleague reveals a possibility that you had not contemplated, in which case you may add the quantity (as H_{n+1} above) and continue. Another is that the truly astonishing has occurred, just as it ought to occasionally, and you continue. A third possibility is that you selected probabilities that were insufficiently dispersed. There is some evidence that the psychological process involved in probability assessment can lead to overconfidence in your knowledge. Sometimes it can easily be corrected by using a long-tailed distribution, such as t with low degrees of freedom, in place of a normal distribution, when the combination of prior and inconsistent likelihood leads to a reasonable compromise but with enhanced dispersion (Lindley, 1983). To repeat the point made in the penultimate paragraph of my paper, we are woefully ignorant about the assessment of probabilities and a concerted research effort in this field is important. Healy is right to draw attention to the psychology of the problem.

Cox raises the nature of the probabilities arising in Mendelian genetics. I would like to reserve the word 'probability' to refer to beliefs. Genetics, and similarly quantum mechanics, use 'chances' which are, as Cox would prefer, related to biological phenomena and arise from exchangeability judgments discussed in Section 14, chance playing the role of the parameter ψ there. The older terminology was direct and inverse probabilities. The distinction becomes useful when you wish to consider the simple concept of your probability of a chance, whereas probability of a probability is, in the philosophy, unsound. I do not understand Cox's remark about calibration. When others raise the issue they ordinarily refer to the long-run frequency, whereas Bayesians live around the present, not long runs, and continually adjust by our beloved theorem, responding in a way that is distinct from the frequentist. This is illustrated by their use of present utility functions rather than error frequencies. His claim that frequentists mesh better than Bayesians with the real world seems wrong to me.

O'Hagan is right in claiming that my discussion of nuisance parameters is deficient and I agree with him that the inclusion of extra quantities, that are not of immediate interest, is fundamental to the assessment of probabilities. It is a case of the larger model being simpler and hence more communicative. However, I do not agree with his comment on the conglomerability example for it is the conditional probabilities that are the tangible concepts. A uniform distribution on the integers only makes sense when it means uniform in any finite set.

Whenever there is a discussion about the Bayesian view, someone is sure to bring out the remark about being 'coherent but wrong' and Nelder does not disappoint. You are never wrong on the evidence

that you have, when expressing your beliefs coherently. To appreciate this, try to give a definition of 'wrong'. Of course additional evidence may demonstrate that you were wrong but Bayesians can deal with that, either by changing the conditions, as when you learn that an event on which you have conditioned is false, or by updating by the theorem. Wrong you may often be with hindsight but even frequentists, or likelihood enthusiasts, have that property also.

I do agree with Dawid that 'Bayesian statistics is fundamentally boring'. A copy of this paper was sent to the person who has the fullest understanding of the subjectivist view of anyone I know (Lad, 1996), and his principal comment was that it was boring. My initial reaction was of disappointment, even fury, but further contemplation showed me that he is right for the reasons that Dawid gives. My only qualification would be that the theory may be boring but the applications are exciting.

Sprott, in his point (e), argues that you should summarize empirical evidence without reference to preconceived ideas and says that this should be done through likelihood statements. Against this I would argue that no-one has succeeded in describing a sensible way of doing this. I dispute Armitage's claim that the 'Fisherian revolution' accomplished this because, although his methods were superb, his justifications were mostly fallacious. Likelihood will not work because of difficulties with nuisance parameters and because of absurdities like that described in Section 13.

An interesting feature of the comments is an omission; there is little reference to the subjectivity advocated in the paper, which surprises me because science is usually described as objective. Indeed Cox, in concluding his advocacy of the eclectic approach, gives a personalistic reason for supporting his view.

> 'I regard the frequentist view as primary, for most if not virtually all the applications with which I happen to have been involved.'

My advocacy of the subjective position is based on reason, subsequently supported by experiences of myself and others.

I conclude on a personal note. When, half a century ago, I began to do serious research in statistics, my object was to put statistics, then almost entirely Fisherian, onto a logical, mathematical basis to unite the many disparate techniques that genius has produced. When this had been done by Savage, in the form that we today call Bayesian, I felt that practice and theory had been united. Kingman's sentence is so apt to what followed.

> 'Perhaps mathematicians select themselves by this desire to reduce chaos to order and only learn by experience that the real world takes its revenge.'

The revenge came later with the advocacy of the likelihood principle by Barnard, and later Birnbaum, so that doubts began to enter, and later still, as the plethora of counter-examples appeared, I realized that Bayes destroyed frequency ideas. Even then I clung to the improper priors and the attempt to be objective, only to have them damaged by the marginalization paradoxes. More recently the subjectivist view has been seen as the best that is currently available and de Finetti appreciated as the great genius of probability. It is therefore easy for me to understand how others find it hard to adopt a personalistic attitude and am therefore grateful to the discussants for the reasoned arguments that they have used, some of which I might have myself used in the past.

References in the comments

Balding, D. J. and Donnelly, P. (1995) Inference in forensic identification (with discussion). *J. R. Statist. Soc.* A, **158**, 21–53.

Bernard, C. (1957) *An Introduction to the Study of Experimental Medicine* (Engl. transl.). New York. Dover Publications.

Cox, D. R. (1978) Foundations of statistical inference: the case for eclecticism (with discussion). *Aust. J. Statist.*, **20**, 43–59.

———(1995) The relation between theory and application in statistics (with discussion). *Test*, **4**, 207–261.

———(1997) The nature of statistical inference. *Nieuw Arch. Wisk.*, **15**, 233–242.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Curnow, R. N. (1999) Unfathomable nature and Government policy. *Statistician*, **48**, 463–476.

Dawid, A. P. and Mortera, J. (1996) Coherent analysis of forensic identification evidence. *J. R. Statist. Soc.* B, **58**, 425–443.

———(1998) Forensic identification with imperfect evidence. *Biometrika*, **85**, 835–849.

Díaz-Francés, E. and Sprott, D. A. (2000) The use of the likelihood function in the analysis of environmental data. *Environmetrics*, **11**, 75–98.

Doll, R. and Hill, A. B. (1950) Smoking and carcinoma of the lung: preliminary report. *Br. Med. J.*, ii, 739–748.
Donnelly, P. and Friedman, R. D. (1999) DNA database searches and the legal consumption of scientific evidence. *Mich. Law Rev.*, **97**, 931–984.
Durbin, J. (1987) Statistics and statistical science. *J. R. Statist. Soc.* A, **150**, 177–191.
Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
———(1993) Bayes and likelihood calculations from confidence intervals. *Biometrika*, **80**, 3–26.
Everitt, B. S. (1999) *Chance Rules: an Informal Guide to Probability, Risk and Statistics*. New York: Springer.
Finney, D. J. (1978) *Statistical Method in Biological Assay*, 3rd edn. London: Griffin.
Gianola, D. (2000) Statistics in animal breeding. *J. Am. Statist. Ass.*, **95**, 296–299.
Healy, M. J. R. (1999) Paradigmes et pragmatisme. *Rev. Epidem. Sant. Publ.*, **47**, 185–189.
Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (with discussion). *Statist. Sci.*, **14**, 382–417.
Lad, F. (1996) *Operational Subjective Statistical Methods*. New York: Wiley.
Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.
———(1971) *Bayesian Statistics: a Review*. Philadelphia: Society for Industrial and Applied Mathematics.
———(1977) A problem in forensic science. *Biometrika*, **64**, 207–213.
———(1978) The Bayesian approach (with discussion). *Scand. J. Statist.*, **5**, 1–26.
———(1982) The use of randomization in inference. *Philos. Sci. Ass.*, **2**, 431–436.
———(1983) Reconciliation of probability distributions. *Ops Res.*, **13**, 866–880.
Lindley, D. V. and Novick, M. R. (1981) The role of exchangeability in inference. *Ann. Statist.*, **9**, 45–58.
Nelder, J. A. (1999) From statistics to statistical science. *Statistician*, **48**, 257–267.
O'Hagan, A. (1988) *Probability: Methods and Measurement*. London: Chapman and Hall.
———(1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Schwartz, D. (1994) *Le Jeu de la Science et du Hasard*. Paris: Flammarion.
Stockmarr, A. (1999) Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics*, **55**, 671–677.
Walley, P. (1996) Inference from multinomial data: learning about a bag of marbles (with discussion). *J. R. Statist. Soc.* B, **58**, 3–57.
Yates, F. (1950) The influence of "Statistical methods for research workers" on the development of the science of statistics. *J. Am. Statist. Ass.*, **46**, 19–34.
Yates, F. and Cochran, W. G. (1938) The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556–580.
Young, G. A. (1994) Bootstrap: more than a stab in the dark (with discussion)? *Statist. Sci.*, **9**, 382–415.