CEDRIC A. B. SMITH

THE ANALOGY BETWEEN DECISION AND INFERENCE

Professor Birnbaum, in his paper (1976), makes a reasonable and valid distinction between 'evidence', i.e. information affecting beliefs, and 'decision', i.e. choice of course of action. He claims that the Neyman–Pearson theory could be looked at in two ways, either as a theory of inference, that is, of evidence, or as a simple kind of decision theory. In fact, he says there are really two distinct 'Neyman–Pearson Theories'. In the first he calls a conclusion obtained from the theory an 'elliptical' or 'evidential' decision, in the second a 'literal' decision. L. J. Savage and D. V. Lindley have presented arguments claiming to show that a reasonable user of the Neyman–Pearson theory should behave as if he was a Bayesian, even if she (or he) calls herself (or himself) a non-Bayesian. [For convenience, from now on we use 'she' to mean 'she or he'.] Professor Birnbaum admits the validity of these arguments in the 'literal' decision case, but argues that they are fallacious in the 'evidential' case.

Before commenting on these points I would like to say that ever since I came to study statistics the Neyman–Pearson theory has impressed me as a very notable intellectual achievement. Since then I have also been greatly impressed by the work of L. J. Savage and D. V. Lindley and A. Birnbaum. (In fact, my first meeting with Savage was at a lecture at which he expounded an argument essentially the same as one quoted in Birnbaum's paper, and at the time it seemed striking and convincing.) I therefore argue here only apologetically and with regrets that the Neyman–Pearson theory is not a reasonable theory of inference, and only in very restricted circumstances is it at all useful as a decision theory, the Lindley–Savage argument is irrelevant, and Birnbaum's objections are based on a misunderstanding. Perhaps it might be as well to add that these rather sweeping condemnations concern the philosopher and theoretician more than the practical statistician. Thus even if (as seems unlikely) all statisticians were instantly converted to the view that significance tests are less than best possible, they would still use them in some situations for

Synthese **36** (1977) 71–85.

a long time to come, because some tests are very simple, quick, and easy to apply, because it will take time to find suitable alternatives, and because significance tests are traditional and widely accepted. Furthermore, most theories of inference or decision view them as 'one person games', in which a single statistician or investigator tries to find what conclusion is best from her own point of view, ignoring other people's interests. In fact, a practising statistician usually works in a social context; she has to convince not only herself, but other investigators and scientists as well, and that may well influence her choice of methods. But, all the same, the most important role played by the Neyman–Pearson theory may have been to introduce a lucid analysis of the problems of inference and decision which was to lead eventually to the more satisfactory approach of L. J. Savage.

1. DECISIONS

The Neyman–Pearson theory is expressed in terms of 'accepting' or 'rejecting' hypotheses. These choices are not, as they stand, clear-cut decisions or courses of action. It has been suggested that to 'accept' a hypothesis could be interpreted to mean 'acting as if it were true', and this does give a rather more precise specification. However, it does not seem entirely satisfactory, since it does not bring out the questions of profits and costs (or advantages and disadvantages) which must be involved in any real decision. Suppose that an experiment tests a fertilizer F for tomatoes. The null hypothesis, H_0, that the fertilizer is ineffective, may be rejected (say) at the 1 per thousand level; that is, the alternative hypothesis H_1 of effectiveness may be 'accepted'. However the practical question at issue will usually be whether to use or not to use the fertilizer. We may perhaps interpret 'acting as if the alternative hypothesis was true' to mean actually using the fertilizer. But whether we do so depends at least on its price and the gain in yield due to its use, as well as perhaps other factors such as climate, availability, and so on, and a proper decision theory must bring these in. The rather simple Neyman–Pearson formulation seems neither appropriately expressed nor adequate to deal with the issues.

The Neyman–Pearson Theory may however be looked on as a step towards the work of A. Wald (1950) and later of L. J. Savage (1954). We

take a 'decision' in favour of a course of action A rather than B to mean the following: if you can choose between A and B (all other relevant circumstances being equal) you will choose A without question. It therefore follows that one cannot both prefer A to B and B to A. From this, and a few similar plausible assumptions, Savage constructed rules which a scheme of preferences might be 'expected' to obey. One might add that these assumptions apply strictly speaking only to that non-existent character, the reasonable and self-consistent human being. Ordinary people may not completely obey his rules. But, if one is trying to work out some kind of theoretical framework for a theory of either inference or decision, it does seem necessary to assume as a minimum self-consistency, even if reality fails to live up to the ideal: otherwise one will have only a self-contradictory confusion. Now Savage showed that those assumptions necessarily imply that the person making the decision will attach to each event E_i a precisely defined (usually subjective) probability p_i and a (usually subjective) value (or 'utility') u_i, so that a person will always choose that course of action providing the maximum expected utility ($\Sigma p_i u_i$). C. A. B. Smith (1961) showed that it was possible to relax the assumptions without greatly affecting the final results. For details, see the original papers. Smith's arguments have been criticized in detail by A. Runnalls (personal communication). But, so far, I am not aware of any serious challenge to the substantive correctness of the theory, which would appear to give a theoretical solution to the one-person decision problem. Note that it involves neither the Neyman–Pearson Theory nor the Lindley–Savage argument in that form, although the discussion does proceed in a rather similar way.

2. INFERENCE

One main difficulty with the use of the Neyman–Pearson Theory is its sharp division of judgment into two categories, 'rejection' and 'acceptance', or 'significance' and 'nonsignificance'. Life is not like that. The results of an experiment on the effect of a fertilizer will vary continuously between the two extremes of showing no effect and virtually conclusively showing a large effect. Having a single threshold between 'nonsignificant' and 'significant' effects is a very arbitrary procedure. Often investigators use several such thresholds, such as the 5%, 1% and .1% levels. These

levels are still arbitrary, even though less so than just a single threshold. Clearly what is wanted is a continuously variable measure of how probable the various hypotheses are, in the light of the data, and the Neyman–Pearson Theory fails to provide this. One must conclude that it is not an appropriate theory of inference.

It is sometimes suggested that we may consider the 'acceptance' of a hypothesis as regarding it as so probable that in practice one takes it for granted. Again life is not like that. What one takes for granted depends on circumstances. There is a bus stop at the end of my road from which buses should run every few minutes to London Airport, taking one hour on the journey. Normally I take it for granted, without thinking about it, that they do; but then, normally, I only need the buses to go to a big shopping centre. If I am actually taking a friend to catch a plane, I no longer take the bus service for granted, but allow at least an extra hour to make sure of being on time.

3. THE LINDLEY–SAVAGE ARGUMENT

The Lindley–Savage argument claims to show that acceptance of the Neyman–Pearson Theory leads inevitably to a Bayesian formulation. If the Neyman–Pearson Theory applies neither to decisions nor to inference, it would seem that the Lindley–Savage argument is irrelevant. What then becomes of Birnbaum's refutation?

It is perilous to attribute motives to others who cannot be asked whether they are being quoted correctly. But my personal impression was that the Lindley–Savage argument was a kind of missionary effort by Savage, saying to the heathen, in this case followers of Neyman and Pearson, 'This will show you that you are already in agreement with me at heart'. It is not really a demonstration from first principles of the correctness of Savage's theory, which can be found elsewhere (Savage, 1954). To understand the point of the Lindley–Savage argument, I suspect that one has to take into account the interpretation of a significance level α as a measure of the intensity of conviction produced by a test. This interpretation, even if implicit rather than explicitly stated, seems a very common part of statisticians' practice. Thus a 5% significance is usually considered as a very mild rejection of a hypothesis, but a 1‰ level almost incontrovertible. More subtly, it is sometimes suggested

that the intensity of conviction should depend on both the significance level α and power $1-\beta$, though it is rarely explained in detail how this is to be done. What Savage is saying effectively "I will show you how to fill this gap. If you think that the intensity of conviction depends only on α and β, and on no other property of the data, then I will show that it must be a function only of some linear function of α and β".

Birnbaum observes that if we have three tests, T_1 with $\alpha = 0, \beta = 0.1$; T_2 with $\alpha = 0.1, \beta = 0$; T_3 with $\alpha = 0.05 = \beta$, then a mixture consisting of the tossing of a coin to decide whether to use T_1 or T_2 will have the same α and β values as T_3. But it will not have the same evidential value, because if in some case we apply T_1 and reject H_0, we know for certain that H_0 is untrue, since $\alpha = 0$ in T_1; if we apply T_2 and accept H_0, we know for certain that it is true, but the unmixed test T_3 can never lead to such certain conclusions. However, Birnbaum's conclusion is not surprising, since it abandons the assumption that the degree of conviction depends only on α and β, since the extra information as to which test is used is also taken into account. Indeed, further consideration suggests that this assumption is not a necessary part of the Neyman–Pearson Theory, which concerns the consequences of using particular values of α and β rather than the degree of belief associated with them. Moreover the assumption is hardly tenable. Two samples may both be classed as 'significant', yet one may be a borderline case, and the other very conclusive. In addition, one's opinions may be considerably influenced by background knowledge, that is by factors additional to α and β, as well as by the observed sample.

4. DISCRIMINATION BETWEEN TWO SIMPLE HYPOTHESES

If we reject the contributions of the other authors referred to above, can we provide a better solution to the problem of inference? There is a well-known difficulty here, in that there can be no absolute and universal standard of correctness in matters or probability. In a statement not involving probability, e.g. that 'Boston lies north of New York', one can test whether it is true or false, and there the matter ends. (In America this statement is true, in England false.) (One may argue that in principle there is a small degree of uncertainty about the truth of any statement, but this uncertainty is negligible for our present purposes.) On the other

hand, if I assert that 'It is very probable that it will rain tomorrow', the occurrence of rain does not conclusively prove this assertion true or false, neither does the absence of rain. The most one can do in such questions is to produce what one feels are reasonable and plausible arguments. If others find them unconvincing, that is that: there is no way of compelling others to agree. What follows can therefore be no more than arguments which seem reasonable and plausible.

It is known that the distribution of a sample conditional on a set of sufficient statistics does not depend on the parameters. (This is effectively the definition of a sufficient statistic.) Hence, once the values of the sufficient statistics are known, it seems plausible that no other property of the sample can provide information relevant to the values of the parameters. (For a distribution which does not depend in any way on a parameter θ cannot throw any light on the value of θ.) But the relative likelihood, i.e. for a sample S the value of [constant $\times \Pr(S|\theta)$] considered as a function of θ, is readily shown to be a sufficient statistic. Hence, in either inference or decision on the basis of a sample, there is no point in considering any property of the sample other than the likelihood.

If we have only a pair of simple hypotheses H_0 and H_1 to compare, this means that the only relevant factor is the likelihood ratio $\lambda = \Pr(S|H_1)/\Pr(S|H_0)$.

This is confirmed by a classic paper by Welch (1939), in which he considers the discrimination between H_0 and H_1 from a number of different points of view, and in every case comes up with the conclusion that it depends only on λ. This is further commented on by Smith (1947). (This means that the degree of conviction produced by the data must be a function of λ and of no other property of the data so long as H_0 and H_1 are the only possible hypotheses.) Note that in fixed-size significance tests λ is not a function of α and β, and there is no fixed relationship between them. (In Wald's sequential method, the test terminates with a value of λ related to α and β.) If we apply a significance test, whether we fixed size or sequential in a mechanical way repeatedly to a long series of different sets of data, the values of α and β concern the overall performance of the series. But the persuasiveness of each individual test depends on the value of λ in that test, and not on α or β.

But suppose that this argument falls on deaf ears. A statistician – let us call her Stella – says, "You may say what you like, but I'm sure my method

of analysing data is better than yours". Have we any grounds for arguing with her?

Stella could be a believer in any one of various methods of statistical analysis – fixed-size significance tests, sequential tests, likelihood, confidence intervals, fiducial intervals, Bayesian intervals, decision procedure, or some less orthodox method, possibly of her own invention. Can we make any general observations, whatever method she prefers? In all these methods with any given observed sample S_s, we associate some 'conclusion' (or 'decision' in the wide sense) C_c. In most cases the method will associate one single, definite conclusion C_c with S_s, but one could also consider the theoretical possibility of a randomized conclusion; that is, after observing S_s we conclude C_c with some probability, R_{sc} say, so that $\Sigma_p R_{sp} = 1$. We return to such randomized conclusions later on: for the present let all conclusions be unrandomized. To simplify the argument, we suppose that there are only a finite (but possibly large) number of possible samples S_s, and only a finite (but possibly large) number of conclusions C_c. This assumption is often untrue, but seems unlikely to lead to seriously unrealistic conclusions, since a continuous variable can be considered as a limit of a variable taking a large number of closely spaced values. One other problem is that in real life, no situation ever quite repeats itself – each set of data is in some way individual and its background different from preceding ones. All the same, if we keep on considering sets of data long enough, we can expect nearly the same kind of situation to be repeated from time to time. By the 'same kind of situation' we mean not that the objects sampled are alike, but that if we are testing, say, two hypotheses H_0, H_1, we have about the same degree of confidence (or prior probability) that H_0 or H_1 is true. It therefore seems not too great a violation of realism to imagine that the position is that we have a large number, N, say, of sets of comparable data (i.e. the 'same kind of situation') in each of which we obtain some sample S_s and then draw some conclusion C_c. We can suppose that, the probability $\Pr(S_s|H_h) = P_{sh}$ obtaining sample S_s conditional on hypothesis H_h is known.

Stella says that, in her opinion, given sample S_s, the conclusion C_c is the 'best' possible. This implies some standard of judgment. This standard may well depend on the method of inference employed. But it seems reasonable that it should be based on the anticipated agreement of

conclusion and reality; for what is the point of coming to a conclusion at all, if it is not hopefully related to reality?

If the object of statistical analysis is to find out the truth, then we might count how often we have achieved a 'correct' conclusion as a measure of the goodness of the method of analysis. Thus, suppose we had rejected hypothesis H_0. Suppose we went to some further investigation, and found out that in fact H_0 was untrue. We would then say that the rejection was 'correct'. Similarly, if we constructed a confidence interval for a parameter θ, and later investigation showed that θ did fall in the interval, we could claim it as 'correct'. However in some more complicated procedures it may be less easy to dichotomize conclusions simply into 'correct' and 'incorrect' ones and even if we do, it is not quite obvious that it is reasonable to regard correct rejections of a hypothesis at the 5% and 1% levels as equally good. So we will suppose more generally that if we have adopted some conclusion C_c, and then later find that hypothesis H_h is true, this gives some 'feeling of satisfaction' (or dissatisfaction) F_{ch}. Obtaining a 'correct' conclusion could be taken as one example of F_{ch}, if one wishes. We do not necessarily suppose that the F_{ch} are measurable in any numerical way, but inasmuch as some conclusions are more satisfactory than others, it is plausible that they should be at least partially ordered. That is, between pairs F_{ch}, F_{di} there may be a relation 'F_{ch} is as satisfactory as or more satisfactory than F_{di}', written $F_{ch} \geqslant F_{di}$ or $F_{di} \leqslant F_{ch}$, and we will take this relation to be transitive,

$$F_{bg} \geqslant F_{ch} \text{ and } F_{ch} \geqslant F_{di} \Rightarrow F_{bg} \geqslant F_{di} .$$

This allows the possibility of incomparable pairs F_{di}, F_{ej}, for which neither $F_{di} \geqslant F_{ej}$ nor $F_{di} \leqslant F_{ej}$ are assorted. However, since we are assuming that these are only a finite number of possibilities, it is not difficult to show that we can if we wish turn a partial ordering (containing incomparable pairs) into a complete ordering by introducing where necessary new relations such as $F_{di} \geqslant F_{ej}$ or $F_{di} \leqslant F_{ej}$, taking care not to violate (1). We comment further about this later, but here let us suppose this done.

We also suppose that if a procedure Π_1 leads to a set of conclusions which are demonstrably less satisfactory than those produced by a procedure Π_2, Stella will be convinced by this that she should not use Π_1.

Consider therefore two possible conclusions C_c, C_d. If we find that

for every hypothesis H_h, $F_{ch} \geqslant F_{dh}$

this implies that Stella will inevitably get just as good results (in her judgment) by using C_c instead of C_d, and possibly better ones. She could therefore omit C_d from the set of decisions, and we will suppose she does so. Hence, if there are only two hypotheses H_0, H_1 under consideration, we will either find

$$(2) \qquad F_{c0} \geqslant F_{d0} \quad \text{and} \quad F_{c1} \leqslant F_{d1}$$

or the same relation with the inequalities reversed. This implies that we can completely order all conclusions, in such a way that

$$C_b \text{ precedes } C_c \text{ precedes } C_d \text{ precedes } \ldots$$

in the ordering means that

$$(3) \qquad \begin{matrix} F_{b0} \geqslant F_{c0} \geqslant F_{d0} \geqslant \ldots \\ F_{b1} \leqslant F_{c1} \leqslant F_{d1} \leqslant \ldots . \end{matrix} \quad \text{and}$$

Now a difficulty in theorizing about statistical inference is that, in practice, history never repeats itself exactly. No problem is exactly like any preceding one, and in particular, even if in each problem there are only two hypotheses H_0 and H_1 considered, in some of them the hypothesis called H_0 may appear more plausible *a priori*, in other H_1. Let us again artificially simplify reality by assuming that there are only a finite (even if large) number of different backgrounds, i.e. different relative prior probabilities or degrees of confidence in H_0 and H_1. Consider a very large series of problems; from them, extract for consideration all those which have some given background, and the same probabilities $\Pr(S_s|H_h) = P_{sh}$ of obtaining a sample of type S_s. If the original amount of data was large enough, we will still have a large amount left in this selection. Suppose that there were N_s samples of size s, and that if we continued to work on them we would find that in N_{s0} of them H_0 was true, and in N_{s1} of them, H_1. If there were altogether n_0, n_1 samples respectively in which H_0, H_1 were true, then by the law of large numbers

$$(4) \qquad N_{s0} : N_{s1} = n_0 P_{s0} : n_1 P_{s1} = n_0 \lambda_s : n_1$$

nearly enough, where $\lambda_s = P_{s0}/P_{s1}$ is the likelihood ratio for sample S_s. Since $N_{s0} + N_{s1} = N_2$, this gives

$$(5) \qquad N_{s0} = N_s n_0 \lambda_s / (n_0 \lambda_s + n_1); \; N_{s1} = N_s n_1 / (n_0 \lambda_s + n_1).$$

Thus, if we assign conclusion C_c to sample S_s, we will find N_{s0} samples giving satisfaction F_{c0} and N_{s1} samples giving satisfaction F_{c1}, where N_{s0} and N_{s1} are given by (5).

Suppose that Stella declares that it is unquestionably best to assign conclusion C_c to samples of type S_s, and conclusion C_d to samples S_t, where C_c precedes C_d in the sense defined by (3). Then I suggest that this is reasonable only if $\lambda_s > \lambda_t$. That is, what conclusion can be assigned to each sample is at least limited by the likelihood ratio; if we found that $\lambda_s \leqslant \lambda_t$ we would accuse Stella of unreasonableness. The argument for this is as follows. Suppose that, for definiteness, $N_s \geqslant N_t$. (If $N_s < N_t$ an analogous argument applies.) From the set of N_s samples of type S_s select at random a subset Σ_s of only N_t samples of this type. The sample in Σ_s will have the same probabilities P_{sh} and the same likelihood ratio λ_s as all samples of type S_s, and if Stella assigns conclusion C_c to all samples of this type, it seems to follow that she will assign it to all in Σ_s. That is, by restricting attention to Σ_s, we see there is no loss of generality in supposing that

$$N_s = N_t . \tag{6}$$

Suppose, if possible that $\lambda_s < \lambda_t$. Then from (5) and (6) it follows that $N_{s0} < N_{t0}$, $N_{s1} > N_{t1}$. Because we assign conclusion C_c to Σ_s, and C_d to S_t, it follows that we have N_{s0} feelings F_{c0}, and N_{t0} feelings F_{d0}. If only we interchanged the conclusions, assigning C_d to Σ_s, C_c to S_t, we would know a greater number of the more satisfactory feelings F_{c0}, and a smaller number of the less satisfactory ones F_{d0}. The same applies to F_{c1}, F_{di}. Thus Stella, despite her protestations, is not doing as well as she might; she would do better to interchange conclusions C_c, C_d. If $\lambda_s = \lambda_t$ the contradiction is less strong: it is simply that, while Stella asserts that interchanging C_c and C_d would be damaging, it will leave in fact the total amount of satisfaction unchanged. The only possibility left without contradiction is that $\lambda_s > \lambda_t$, as stated above.

5. INFERENCE ABOUT MORE THAN TWO HYPOTHESES

The argument of the last section suggests a relation between the likelihood ratio λ_s of a sample S_s and the appropriate conclusion C_c to be drawn from it. It assumes that Stella has already chosen her favorite

method of statistical inference, which we examine, and possibly criticize. It uses several devices, such as the replacement of a partial order by a total order, and the use of the law of large numbers, which seem not too unreasonable, but might be challenged. And as it stands, it deals only with the case of two hypotheses, H_0 and H_1. Can we do better than this?

In the corresponding problem in one-person decision theory, Savage (1954) and Smith (1961, 1971) proceed by considering randomized decisions and mixtures of these together with some mild and apparently reasonable assumptions. For Savage, preferences are completely ordered, and he deduces that each possible outcome i, or result of a decision, has a certain numerical probability p_i and a certain numerical value or measure of advantage or 'utility' u_i. The utility of a decision is the expectation of the utility taken over all its possible outcomes, i.e. $\Sigma p_i u_i$. The decision actually made will be the one with the greatest utility. It also follows that an actual decision need never be randomized, i.e. the idea of a randomized decision is a mathematical tool to support the theoretical arguments, and it can be discarded at the end. It also follows from Savage's theory that if we have a set of hypotheses H_h, their final (or posterior) probabilities following the drawing of a sample S_s will be derived from the initial probabilities by the use of Bayes' Theorem. In fact, they obey all the usual laws of the calculus of probabilities. There may be any (finite) number of hypotheses H_h involved. Smith (1961) suggests that the assumption of a complete ordering of preferences would seem too strong to most people, and shows that a very analogous theory can be developed using the weaker assumption of partial ordering. This theory does not require that the partial ordering should be arbitrarily replaced by a complete ordering, and neither theory requires that we should have a large collection of samples of data and make use of the law of large numbers.

The difficulty in applying this type of argument to inference problems has always been the apparently fundamental difference between a decision problem, in which one tries to get the maximum advantage, and an inference problem, in which one tries to find the truth. However, a more careful examination suggests that the differences are not so great as might at first be thought. What a 'material advantage' is depends on psychological judgment, as well as on physical factors. Thus most people, on retiring from their jobs, would look on a small present from their fellow-workers

as a useful and advantageous gift as well as a token of esteem. But, even while this paper is being written, I have met a friend who on his last day at work went instead round to his colleagues giving *them* presents, and quietly departed. Similarly, as we have already pointed out, strictly speaking a statistical conclusion is almost never either proved or disproved by subsequent observational evidence. We may say that we 'reject' a hypothesis H_0, or 'assert' that a parameter lies in a certain interval, but such 'rejections' or 'assertions' are almost always uncertain; they indicate the way the evidence seems to point more or less strongly, but always allowing the possibility of the opposite. Thus the relationship between a conclusion C_c and a hypothesis H_h is best considered as 'degree of agreement', or 'feeling of satisfaction' F_{ch} if the hypothesis is found to be true, or some similar expression, and not black-and-white truth-or-falsity. Thus the inference problem, like the decision problem, relates also to matters of psychological judgment. Although they are different problems, they both, strictly speaking, involve subjective preferences. Provided that one can supply the necessary tools, namely a 'randomized conclusion' and a 'probability mixture of conclusions', the arguments applied in decision theory will also apply similarly to inference, with appropriate modification of terminology. Since the conclusion arrived at in any particular case depends on the maximization of 'expected satisfaction', which depends on the final (posterior) probabilities, we may accordingly deduce (by analogy with the arguments of decision theory) that the fundamental problem of statistical inference is the calculation of these probabilities, and that is in principle solved by Bayes' Theorem provided that we can find reasonable initial (prior) probabilities.

We must therefore answer the question: can we have 'randomized conclusions', i.e. in some given situation can we choose at will to come to conclusion C_c with some specified probability R_c? There is no problem here when randomized decisions are concerned, but with conclusions an investigator may say: "I have no choice. The evidence leads me to conclusion C_c, and if I said I could support a different conclusion C_d, I would be dishonest". The difficulty is particularly acute if the conclusion C_c represents the 'degree of belief' induced by the evidence, which is something which cannot be varied at will. However, perhaps we can say to the participant: "At least imagine that you could be mistaken, that you

could conceivably reach a different conclusion, and see what the consequences would be". In this way one could at least hypothetically consider a randomized conclusion, (and such hypothetical consideration seems all that is necessary to justify the mathematical argument).

If we have two randomized conclusions, $\Sigma R'_c C_c$ and $\Sigma R''_c C_c$, then for any non-negative numbers k', k'' such that $k'+k'=1$, the 'mixture' $\Sigma(k'R'_c+k''R''_c)C_c$ will be a randomized conclusion. If by a 'conclusion' we mean some statement of the form 'θ is significantly different from 0' or 'θ lies in the interval $\theta_1 \leqslant \theta \leqslant \theta_2$' rather than an actual subjective degree of belief, there certainly seems no difficulty in forming such a mixture.

But what about Birnbaum's objection to probability mixtures of conclusions? We might paraphrase it somewhat as follows. "In the University of Exbridge, Men and Women Academic Staff have separate Common Rooms. There is a Corridor A with a fork at the end, the left leading only to the Ladies' Common Room, the right to the rest of the College. There is another Corridor B, with a fork leading right only to the Mens' Common Room, left to the rest of the College. We see a lecturer walking along, whose sex is not obvious, and wish to check. If we observe her/him walking down Corridor A, and see her turn left, there is no doubt that she is female; if she/he turns right, the matter is left in doubt, although the chances of maleness are increased. Similarly (with sex interchanges) for Corrdior B. Or one can adopt a mixed strategy, by tossing a coin, and watch the lecturer go down Corridor A if the coin falls heads, and B if tails." Now if the conclusion is reported only in the form of either 'she/he turns left' or 'she/he turns right', without mentioning which corridor is involved, it would obviously be better to use an unmixed strategy in which the particular corridor had been specified beforehand. However, if the observations were divided into four separate cases specifying both the corridor (A or B) and the direction of turn (left or right), then the objection to use of a probability mixture would seem to disappear so long as these four distinct cases are not confused with one another. In any case, we are here speaking of a mixed *strategy* rather than a mixed conclusion. Thus Birnbaum's example does not seem to constitute an insurmountable obstacle.

Granted that these suggestions are valid, we can construct a theory of inference or evidence strictly parallel to that of decisions. This would be hardly surprising, since how one acts often shows vividly what one

believes, and a considerable divergence between theories of decision and inference would be difficult to explain.

6. LIMITATIONS OF THE THEORY

Ths theory outlined above has two important limitations. In the first place, we suppose that the numbers of possible samples, hypotheses and conclusions are all finite, whereas, for example, a statistical parameter could be thought of as taking any real value within some interval. One would suspect that, at least in cases likely to arise in practice, the theory would carry over in the obvious way, but I know of no rigorous demonstration, I also suppose that the hypotheses H_h could be found true or false (for practical purposes, even with no absolute certainty) by sufficient additional work. Thus we do not consider the test of scientific theories, such as General Relativity, if we suppose they can only be refuted, never verified with certainty. This limitation is important. But it should be noted that if, for example, to test Einstein's theory we look at measurements of the deflection of light by the sun, what we immediately estimate is the actual deflection (with the experimental error removed). This is a quantity which could be estimated in principle with arbitrary accuracy, by gathering sufficient data: so we have a valid statistical estimation problem here, quite apart from any light it may throw on the correctness of General Relativity.

I am also aware that many details would have to be filled in carefully in the discussion in this paper, before it could be considered rigorous. For example, we have shown that if Stella makes certain statements, she is acting 'unreasonably'. But in order to justify such an objection, one would also have to prove that it was possible for her to act 'reasonably'. I present the argument, not in a complete, polished and final version, but in a preliminary and imperfect state, to stimulate discussion of whether it has major defects of principle.

University College, London

BIBLIOGRAPHY

Birnbaum, A., 1977, 'The Neyman–Pearson Theory as Decision Theory, and as Inference

Theory; with a Criticism of the Lindley–Savage Argument for Bayesian Theory', *Synthese*, this issue, pp. 19–49.

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

Smith, C. A. B., 1947, 'Some Examples of Discrimination', *Annals of Eugenics* **13**, 272–282.

Smith, C. A. B., 1961, 'Consistency in Scientific Inference and Decision', *Journal of the Royal Statistical Society B* **23**, 1–37.

Smith, C. A. B., 1971, 'Simple Game Theory and Its Applications', *Bulletin of the Institute of Mathematics and Its Applications* **7**, 352–357.

Welch, B. L., 1939, 'Note on Discrimination Functions', *Biometrika* **31**, 218–220.