D. V. LINDLEY

THE DISTINCTION BETWEEN INFERENCE AND DECISION

This paper is a commentary on that by Birnbaum (1977). The first part states the Bayesian position on the points raised by him. The second part discusses his position in the light of the Bayesian approach.

1. THE BAYESIAN ARGUMENT

In the case of data, x, dependent on a parameter, θ, possibly vector-valued, through a known probability density $p(x|\theta)$, *inference* is effected by stating the distribution of θ given the data, $p(\theta|x)$, obtained from the original distribution of θ, $p(\theta)$, by Bayes' formula,

$$(1) \qquad p(\theta|x) \propto p(x|\theta)p(\theta).$$

Decision may then be effected by introducing for each possible decision d, a utility function $u(d, \theta)$ – the utility to the decision maker of d were he to be informed that the parameter had value θ – and then selecting that decision of maximum expected utility:

$$(2) \qquad \max_d \int u(d, \theta)p(\theta|x)\,\mathrm{d}\theta.$$

The inference and decision processes are therefore different and solved by different methods – Bayes' theorem and maximization of expected utility respectively. The former, in addition to the known $p(x|\theta)$, requires only $p(\theta)$ and the observed data values. The latter has additional ingredients in the class of available decisions and the utility function. Whilst it is possible to make inferences without considering decisions, the implementation of decision-making requires an earlier calculation of the appropriate inference, $p(\theta|x)$. Notice, however, the important point, that *every* decision problem which depends on θ requires the *same* inferential statement in order to evaluate the expected value. This is why it is useful to separate inference from decision because inference can be carried out

Synthese **36** (1977) 51–58. *All Rights Reserved.*

without decisions and yet it is the only aspect of the data relevant to any decision problem to which the data can make a contribution. Alternatively expressed we can refer to the *evidence* $p(\theta|x)$ about θ given the data, and then consider *behaviour* in the choice of an act, or decision.

There are essentially two ways of justifying the position just described. The first is pragmatic: it works. When applied to situations arising in practice it gives sensible results. The second justification is more theoretical. It starts from a consideration of the problem of decision-making under uncertainty – how should d be chosen when one is uncertain about θ. Certain requirements for reasonable decision-making are then laid down. For example, if d is preferred to d^1 when θ is true, and also when θ^1 is true, then d should still be preferred to d^1 when one is uncertain whether the parameter takes the value θ or θ^1. (This is often referred to as the 'sure-thing' principle.) These requirements are taken as axioms from which Bayes theorem and the principle of maximizing expected utility can be deduced. It is now usual to refer to the requirements as axioms of *coherence*, because they state, as in the 'sure-thing' principle, how one judgement coheres with another. The Bayesian argument might better be termed the coherent argument. Recognition of the philosophical position that inferential statements are made because of their potential value in action completes the justification for the inferential argument with Bayes' theorem. In repetition of what has been said above, the inference requires no decision consideration, but is valid for any relevant decision.

Certain aspects of the development need emphasizing in order to apply it to Birnbaum's discussion. Notice that the only contribution of the data to the inference – and therefore to decision – is through the term $p(x|\theta)$ in (1). Furthermore in this expression x will assume the numeric values observed when the data is to hand, whereas we will need to contemplate *all* values of θ. The above convenient, but rather incomplete, notation would be better written $p(x|\cdot)$ emphasizing that it is considered as a function of its second argument for the fixed, observed value of its first, namely x. This is the likelihood function; a function of θ for fixed x. An immediate deduction from the coherent position is the likelihood principle that says that all evidence supplied by data x is contained in the likelihood $p(x|\cdot)$.

Birnbaum considers the case of two simple hypotheses H_1 and H_2, or in our language, two parameter values θ_1 and θ_2. In this situation (1) may be written

$$\text{(3)} \qquad \frac{p(\theta_1|x)}{p(\theta_2|x)} = \left\{\frac{p(x|\theta_1)}{p(x|\theta_2)}\right\} \frac{p(\theta_1)}{p(\theta_2)},$$

avoiding the unstated constant of proportionality in (1). The expression in braces is called the likelihood ratio and provides the only contribution of the data to the inference and decision problems. The Neyman-Pearson errors of the two kinds, α and β, require more than the likelihood ratio, even though that quantity plays a dominant role in their theory, because α, for example, is $\int_{R_1} p(x|\theta_1)\,\mathrm{d}x$ where R_1 is the set of x-values that will lead to rejection of H_1. Its calculation therefore requires $p(\cdot\,|\theta_1)$, a function of the data, and not merely that function evaluated at the observed data. It is for this reason that the Neyman-Pearson argument is technically incoherent.

There is one aspect of the Bayesian view which is often criticized as being its major weakness: in fact it arises from its great strength, coherence. This is the introduction of the distribution, $p(\theta)$, of θ in advance of the data – the so-called 'prior' distribution. Its introduction means that the inference, (1), requires in addition to the likelihood function a specification of $p(\theta)$. In particular, although two people often agree on the form of $p(x|\theta)$ they may well disagree about $p(\theta)$ and therefore their inferences (in the above sense) will be different. Alternatively expressed, the coherent view is unable to let the data 'speak for themselves' but requires extraneous judgements about θ. To see why this happens it is only necessary to recognise that coherence is concerned with how judgements or actions 'fit together' – or cohere – in particular with how views about θ before the data are available cohere with views after it has been obtained (and with the relationship between x and θ expressed through $p(x|\theta)$). The only way to achieve coherence is through Bayes theorem, (1). Look at it the other way round: any other inference form, such as one that lets the data 'speak for themselves', may be incoherent. Suppose, in particular, that the data came in two sets x_1 and x_2, with $x = (x_1, x_2)$, then the separate statements derived from x_1 and x_2 will not cohere with the statement from x unless Bayes theorem is used. Consequently the apparently unwanted $p(\theta)$ is an advantage in enabling coherence to be achieved.

There is a school of thought that says that inference should stop at the likelihood function. This works well when θ is a real number – and, in particular, in Birnbaum's case where θ only takes two values – but fails in the presence of nuisance parameters. Suppose $\theta = (\theta_1, \theta_2)$ where θ_1 is the only value of interest, θ_2 being present in $p(x|\theta) = p(x|\theta_1, \theta_2)$ but being a 'nuisance'. An example is where x is the measurement of θ_1 with precision θ_2. There the likelihood function involves an irrelevant θ_2. This can be removed in the coherent approach using

$$p(\theta_1|x) = \int p(\theta_1, \theta_2|x)\,\mathrm{d}\theta_2,$$

the marginal distribution of θ_1, but no such device is available in the likelihood method save in special cases where the likelihood function factorizes in an appropriate way.

We have seen that the coherent approach introduces $p(\theta)$ as well as $p(x|\cdot)$ for observed data, x. It is often said that other methods do not introduce extraneous elements. In fact this is not so. Consider the Neyman-Pearson approach with its error-rate α. Its calculation (see above) involves an integration and therefore other data values besides that observed. What are these other values? Suppose we have in 12 trials observed 9 successes and 3 failures – which we write (9, 3) – are the other values (10, 2), (11, 1) etc., where the total number of trials are fixed at 12; or do they include (10, 3) because the experimenter might have had time to observe a thirteenth trial? There are occasions, as with sampling-inspection schemes, where the other possibilities are clear-cut, but with most inferential situations the alternative values are far from clear. It is not uncommon to find the sample size, in our example, 12, fixed without any serious justification being offered. Any argument using error-rates introduces extraneous elements, unobserved data points, so that the coherent viewpoint is no greater offender in using $p(\theta)$.

2. BIRNBAUM'S DISCUSSION

It should be clear from the first section of this paper that Birnbaum's view is different from the coherent one and his paper is a forceful argument against coherence in inference – or as an evidential tool – though not in decision-making – or as a guide to behaviour. In particular he rejects a possible axiom which he calls assumption II*.

(I agree with Smith (1977) in his comments that the Lindley–Savage argument is irrelevant. The general approach to coherence is much more useful and does not begin by begging the question as to whether the error-rates are appropriate. It was developed by us in 1955 partly in a successful attempt by Savage to demonstrate that some ideas of mine were unnecessarily complicated, and partly in reaction to some ideas due I think to Lehmann, though I am unable to locate the reference, which referred to indifference curves in the (α, β)-plane. The argument shows that these 'curves' must form a set of parallel straight lines.)

Let me try to convince you that the description of evidence in the form

(4) (reject H_1 for H_2, α, β)

(§ 3) is unsatisfactory. In the first example of § 8 he says "In such situations I particularly value the guarantee, which is provided by use of (0.05, 0.05), that strong evidence will be obtained (either supporting H_1 against H_2, or supporting H_2 against H_1)." Suppose that on H_1 x is uniformly distributed in the interval (0, 1), and on H_2 it is uniform in (0.9, 1.9). The densities $p(x|H_1)$ and $p(x|H_2)$ are shown in Figure 1. The rule in which we reject H_1 for H_2 if $x > 0.95$, and otherwise reject H_2 for H_1, has $\alpha = \beta = 0.05$ which he values. But consider what happens if $x = 0.99$, say (any value between 0.9 and 1.0 will exemplify the difficulty). This leads to rejection of H_1 and therefore good evidence on the criterion suggested. But is this reasonable, for $x = 0.99$ is just as likely under H_1, as it is under H_2, both probability densities being unity. Is it not intuitively sensible to think that this datum adds nothing to our knowledge of whether H_1 or H_2 obtains? Contrast this with what would happen were $x = 1.01$. According to the recipe we would still (reject H_1 for H_2, 0.05, 0.05) yet now we would have a value which has zero probability density under H_1 – roughly it is impossible – yet still has unit density under H_2. This value is surely very strong evidence in favour of H_2. Consequently I argue that Birnbaum is wrong in saying "strong evidence *will* be obtained" (my italics) – on the contrary, the evidence might, as with $x = 0.99$, be nil.

Notice in the likelihood ratio approach, (3), for $x = 0.99$ we have a ratio of 1 and the probabilities for H_1 and H_2 are unaltered; whereas for

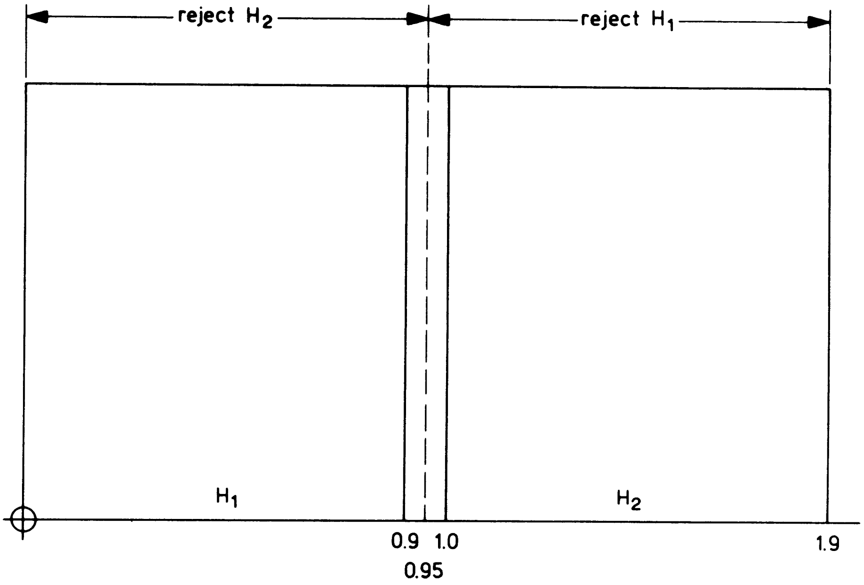

Fig. 1

$x = 1.01$ we have a ratio of 0 and $p(H_1|x = 1.01)$ is zero. This argument therefore agrees with that put forward in the last paragraph.

Another way of seeing that the form (4) is unsound is to remark that the values of α and β are available in advance of the data. Once the class of possible x-values is decided upon (the extraneous element mentioned above) integrations provide the two error-rates. Consequently the only way in which the data influence (4) is whether they lead to reject H_1 or to reject H_2. In other words the evidence – in the sense of knowledge affected by the actual data values – is dichotomous. Contrast this with the richness provided by the likelihood ratio which can typically take values in a continuous range, often from zero to infinity.

Birnbaum's rejection of the mixture axiom (his assumption II*) has important consequences for much of conventional statistical thinking. To illustrate this consider a modification, to discrete values, of the example taken above. Here x takes the values 0, 1 and 2; under H_1 with probabilities 0.9, 0.1 and 0.0, under H_2 with probabilities 0.0, 0.1 and 0.9. By rejecting H_1 whenever $x \geqslant 1$ we get (α, β) – values of (0.1, 0.0): by rejecting H_1 whenever $x \geqslant 2$ we get (0.0, 0.1). In the same example in his paper he suggests the preference pattern

$$(0.05, 0.05) > (0.1, 0) \sim (0, 0.1),$$

so that he is indifferent between the two rules just suggested. No rule is available which gives the preferred (0.05, 0.05). But consider observing, in addition to the datum x, the result of tossing a fair coin. Let us then reject H_1 if $x = 2$, or if $x = 1$ and the coin shows 'heads'. This gives the values (0.05, 0.05) and consequently Birnbaum prefers evidence that uses the toss of a coin whose result is irrelevant to H_1 or H_2 to that which does not. This I find surprising. Furthermore in a general situation with data x and parameter θ let us observe any quantity having a known distribution which does *not* depend on θ. Denotе this by y, then y is called *ancillary*. (Sometimes additional requirements are placed on it before it is given this name but I do not think these are relevant here.) It is usual in conventional statistical practice to argue conditional on the value of any ancillary yet Birnbaum's preferences would deny this. This convention is well supported by the coherent viewpoint since $p(x, y|\theta) = p(y|\theta)p(x|y, \theta)$ and if $p(y|\theta)$ does not depend on θ we have

$$\begin{aligned} p(\theta|x, y) &\propto p(x, y|\theta)p(\theta) \\ &\propto p(x|y, \theta)p(\theta) \end{aligned}$$

using only the conditional distribution.

I conclude with a few miscellaneous points arising from Birnbaum's article. In § 2 he asks whether it is appropriate to think of estimation as a decision problem. Notice that in the coherent framework the estimation problem as such disappears. The answer to every conceivable inference is contained in $p(\theta|x)$ and the question of whether to estimate using the mean or the median does not arise. The only Bayesian sense of estimation lies in the description of $p(\theta|x)$ for someone unable to live with a function $p(\cdot|x)$. Thus we might discuss whether the mean or the median is a better description of the location of the distribution.

He points out that a common application of Neyman-Pearson theory is in industrial sampling inspection. It is relevant to remark that even in this decision environment that theory can give incoherent answers. For example the decisions in a scheme of size n_1 may not cohere with those for one of size n_2. It is interesting to speculate why Neyman and Pearson, and later Wald, never considered the notion of coherence, even in the narrower decision context.

In the footnote to § 7 a concept (*P*) is formulated. It is worth exploring the relation of it to the coherent attitude. Suppose strong evidence against H_1 is interpreted as $p(H_2|x)/p(H_1|x) > k$ where k is some large number, say 100. By (3) this means

$$p(x|H_2) > p(x|H_1)k'$$

where $k' = kp(H_1)/p(H_2)$. Integrating over all x values for which this is true we have $(1-\beta) > \alpha k'$ or $\alpha < (1-\beta)/k' < 1/k'$, a small number unless the prior odds on H_2 are large. This agrees with (*P*). (The argument is familiar in Wald's theory of sequential tests.)

In summary, my view is that Birnbaum's criticism of the mixing assumption is not convincing even in the inferential context and that the meaning he attaches to evidence is not as satisfactory as that based on the likelihood principle and the use of Bayes' theorem.

Department of Statistics and Computer Science,
University College, London

REFERENCES

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.

Smith, C. A. B.: 1977, 'The Analogy between Decision and Inference', *Synthese*, this issue, pp. 71–85.