HENRY E. KYBURG, JR.

DECISIONS, CONCLUSIONS, AND UTILITIES

1. Both the founders and theorists of Bayesian approaches to statistical inference, and the founders and theorists of the 'standard' Neyman–Pearson approach to statistical inference deliberately, explicitly, and self-consciously regarded their theories as theories of rational *behavior*. To take two paradigm personalities: Neyman has become ever more insistant that he is, and has been, concerned with *inductive behavior*; Savage from the outset has taken as the very basis of his theory the principle that the rational agent tries to act so as to maximize his expected utility. (The great exception among the modern giants was Fisher, who denigrated subjective beliefs as irrelevant to science, and statistical decision theory as 'statistics for shopkeepers'.) It is thus hardly surprising that the assumptions of both major schools of statistical inference are more amenable to the behavioral interpretation than to the evidential interpretation of statistical tests.

But the matter is not quite so simple as it may seem. There is a way of bringing together behavioral and evidential concepts. It is not simply a matter of saying that 'accepting' a hypothesis is performing an action; it is more complicated than that. Making a statistical inference can be plausibly construed as an action if we take the *values* of believing hypotheses, suspending judgment, and rejecting hypotheses quite seriously. That is: if we take epistemic utilities seriously. The best known proponent of epistemic utility is Isaac Levi, and he develops a rather full system of inductive inference in his book, *Gambling with Truth*, whose very title suggests how the consideration of epistemic utilities can be used to connect behavioralistic statistical theory with the evidential concerns of scientists and philosophers.

Epistemic utilities are no different in principle from any other utilities. There is a certain value to accepting a hypothesis H_1 when it is true, a certain disvalue about accepting it when it is false. By their very nature, utilities are the sorts of things whose expectations are also utilities. To put

Synthese **36** (1977) 87–96.

the matter more plainly, if the value of accepting H_1 when it is true is U_1, and the value of failing to accept it when it is true is U_2, and a certain decision function d gives us a chance of accepting H_1 of p, and of rejecting H_1 of $1-p$ then when H_1 is true, the value of d is $p \cdot U_1$ plus $(1-p) \cdot U_2$. To deny that the expectations of epistemic utilities are epistemic utilities is to deny that they are utilities at all. This is an extreme move. [But it has been made. W. K. Goosens, in an unpublished paper, 'A Critique of Epistemic Utility', has argued against epistemic utilities.] Let us explore the conservative route first, and assume that epistemic utilities make sense, and that they are utilities in the sense that their probabilistic expectations are also epistemic utilities.

2. Let us consider the simplest sort of situation, as does Birnbaum, in which we have to decide between two hypotheses H_1 and H_2. Let us suppose that the utility of accepting H_1 when it is true is U_1, of failing to accept H_1 when it is true is V_1, and that U_2 and V_2 are the corresponding utilities for H_2. A decision function D is a function from possible items of evidence to decisions d_1^* and d_2^* as characterized by Birnbaum. That is, given the evidence e, we either have $D(e) = d_1^* =$ (reject H_1 for H_2, α, β) or $D(e) = d_2^* =$ (reject H_2 for H_1, α, β), where α and β are, respectively, the probability of falsely rejecting H_1 and the probability of falsely rejecting H_2. Given any decision function D, we can easily enough compute its expected utility under H_1 and its expected utility under H_2: $E_1(D) = (1-\alpha)U_1 + \alpha V_1$ and $E_2(D) = (1-\beta)U_2 + \beta V_2$. In considering ordinary statistical hypotheses, it seems unlikely that there is any piece of information – any hypothesis – which is infinitely valuable, or any mistake that is infinitely disasterous. If this is so, we can simplify things by considering losses rather than utilities and normalizing them to the interval (0, 1). And we might as well compute $\alpha^* =$ the expected loss of D under the hypothesis H_1, and $\beta^* =$ the expected loss of D under the hypothesis H_2, and then characterize our cases of statistical evidence as $d_1^u =$ (reject H_1 for H_2, α^*, β^*) and $d_2^u =$ (reject H_2 for H_1, α^*, β^*).

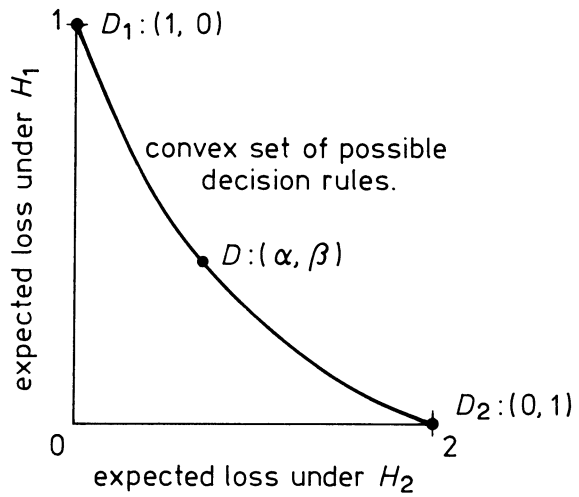
The cases of statistical evidence are now specified in terms of losses. This is in fact more appropriate than the specification in terms of probabilities, both for the classical decision theoretic approach and for the Bayesian approach. The radio manufacturer will make his decision to sell or not to sell according to the losses involved in the two cases when he

follows a Wald decision program; he will make his decision to sell or not to sell by maximizing his expected utility if he follows a Bayesian program. If we can speak of accepting and rejecting hypotheses as actions by means of evaluating them in terms of epistemic utilities, then the same utility considerations will enter into the purely scientific statistical problem. (Of course the utilities are different: one is in dollars, and the other is in cogs – the units of cognition.)

It will help to keep matters straight – and to form a connection with Birnbaum's arguments – if we construct a picture of what is going on. Let the vertical axis of our picture represent the losses under hypothesis H_1, and the horizontal axis represent the losses under hypothesis H_2. We have already stipulated that these losses are to be normalized to the unit interval, which will both make it easy to confuse them with probabilities (a confusion which must be resisted) and also easy to compare Birnbaum's formulation with mine.

The decision rule D will be represented in the unit square by the point (α^*, β^*). Since Birnbaum's α represents the probability of *falsely* rejecting H_1 for H_2, we shall construe α^* as the expected loss of following D when H_1 is true. A rule D_1 which instructed us to reject H_1 for H_2 under all circumstances would have maximum expected loss under H_1, minimum expected loss under H_2, and would thus be represented by the point (1, 0). Similarly a rule D_2 which instructed us always to reject H_2 for H_1 would be represented by the point (0, 1). Let $\mathscr{D}$ be the set of decision rules. This set is convex, since if D_i, characterized by (α_i, β_i) and D_j, characterized by (α_j, β_j) are in the set, so will D_k, characterized by $(kd_i + (1-k)\alpha_j, k\beta_i + (1-k)\beta_j)$ be in the set. This is a direct consequence of dealing with expected utilities. If α_i is one expected utility and α_j another, a probabilistic mixture in the ratio $k:(1-k)$ will have the expected utility $k\alpha_i + (1-k)\alpha_j$.

In this framework, where we are dealing with expected utilities, the Lindley–Savage argument goes through with no difficulty. Assumption II, highlighted by Birnbaum, automatically holds, because utilities just *are* that sort of thing. If I am indifferent between the *utilities* of two alternatives, I will also be indifferent between the utilities of any two probabilistic combinations of them. It will follow that the indifference classes in the unit square will turn out to be parallel lines of a given slope. This slope will be interpreted, by Lindley and Savage, as reflecting the ratio of the prior



probability of H_2 to the prior probability of H_1. The preferred decision rule will correspond to the point (or those points) of the convex set of possible decision rules that rests on an indifference curve which is a supporting line for the set of points representing possible decisions.

Of course, as Birnbaum points out, there are lots of other ways of choosing a decision rule. We may fix on an acceptable level of α^* and then choose that decision rule that minimizes the expected loss β^*, or vice versa. We may adopt a minimax regret principle, which will be represented in our picture by a line of positive slope through the origin: the point of intersection of this line with the boundary of the decision set will be the chosen decision. Or we may adopt a straight minimax principle, which will be represented by a line parallel to the minimax regret line, but perhaps not passing through the origin. (In either case, the slope is determined by the actual, un-normalized utilities involved.)

An admissible decision rule is a decision rule on the boundary on the convex set of decision rules. Every principle I have mentioned contains enough parameters (the prior probabilities, the utilities involved) so that, given any admissible decision rule, given any principle for selecting among decision rules, the parameters can be adjusted in such a way that that principle will pick out that decision rule as 'correct'. In short, if we regard these parameters as adjustable, none of these principles gives us any guidance at all. They are all vacuous. They are vacuous in the

straight-forward down-to-earth shopkeeper case as well as in the pure scientific evidential case.

But this just shows, perhaps, that we cannot regard the parameters as adjustable. If we are given fixed values for the parameters – fixed prior probabilities, fixed losses, and so on – then the application of one of these principles will indeed pick out a decision rule. But now we have a different problem: each of these principles may pick out a *different* decision rule as appropriate! And this is true whether we are considering a behavioral interpretation of statistical evidence with utilities measured in dollars and cents, or an evidential interpretation of statistical evidence with utilities measured in whatever units are appropriate for epistemic utilities.

I have offered elsewhere [*The Logical Foundations of Statistical Inference*, Reidel, 1974] a partial resolution of this problem in terms of interval valued epistemological probabilities. If we suppose that prior evidence rationally imposes constraints on the prior probabilities of H_1 and H_2, then the indifference classes in the unit square may, so far as rational constraints are concerned, be composed of families of lines of slopes bounded by slopes corresponding to the upper and lower probability bounds on the hypotheses. The constraints on the prior probabilities (which on my view are objective, determined by the prior evidence, and not merely a matter of subjective opinion) thus constrain the acceptable decision rules to lie on a proper part of the boundary of the decision space. These constraints are operative whether we are concerned with an evidential interpretation of statistical data, or a properly behavioral interpretation.

In general, of course, there will still be a number of acceptable decision rules. I argued that the constraints imposed by the prior probability interval were as far as logic or rationality could guide us. Beyond this point, when we are concerned with the properly behavioral interpretation of statistical evidence, we may be guided by acceptance level principles, minimax principles, minimax regret principles, or variations on them. What is appropriate will depend on our interests and concerns above and beyond the concern to maximize our expected utility.

Will the same variety of considerations apply to the evidential interpretation of statistical evidence? Is there some way in which we can account for Birnbaum's examples by bringing to bear considerations of the same sort we bring to bear when we need to choose among decision rules which

are all acceptable from the point of view of maximizing expected epistemological utility?

3. The answer is unclear, on the assumption that epistemic utilities make sense. But rather than try to clarify possible answers to these questions, it is now time to look more closely at the assumption on which the questions are based. As Birnbaum remarks, both the Bayesian approach and the approach derived from Wald make the assumption that we are (and should be) indifferent between probabilistic mixtures of decision functions between which we are indifferent. These decision rules are characterized in terms of their error values. It might be thought that the introduction of epistemic utilities could circumvent the difficulties which Birnbaum raises for the assumption in question. The answer, however, is clearly negative. We may consider the epistemic loss involved in falsely rejecting the hypothesis H_1 (or H_2) to be one cog. The expected loss, then, under each hypothesis, will be numerically identical to the error probability. And the example in which, seeking at least the possibility of certainty, we are indifferent between the decision rules characterized by $(0, 0.1)$ and $(0.1, 0)$ but prefer either one to $(0.05, 0.05)$ is just as persuasive as ever. But now the import of the example is that these numbers *cannot* be interpreted as expected utilities. If there is anything that can be said about expected utilities at all, it is that they can be probabilistically combined. Even if we should want, under some circumstances, to reject the principle that an expected utility – i.e., a utility multiplied by a probability – is a utility, that would do us no good here, for the numbers we are combining are *already* expected utilities.

What I take to be the upshot of Birnbaum's persuasive examples, then, is not merely that neither Bayesian nor classical interpretations of statistical evidence can be construed evidentially, but that no notion of epistemic utility can be used to bring the use of statistics in science into line with the use of statistics in shopkeeping. But it is not clear that there is any motivation for wanting to construe the scientific use of statistics along the same lines as the behavioristic use of statistics, apart from the fact that professional statisticians are trained to use a certain bundle of techniques and principles many of which were developed for the behavioristic acceptance-sampling situation. I rather strongly suspect that these circumstances have not been altogether happy ones for the development of

the best statistical treatment of data in an evidential context. It may well be, in fact, that the machinery for 'choosing between two statistical hypotheses' is simply inappropriate to the analysis of statistical data for the sake of developing or refining genetic theory. Or, at any rate, it may be that the appropriateness of that 'conceptual framework' is not so general as has been thought, and depends rather more heavily than has been thought on details of the body of knowledge that the geneticist brings to his problem.

Even assuming the appropriateness of the hypothesis-testing formulation, however, the background information comes to play an important role. Consider the example Birnbaum offers: We are to consider two outcomes which are of equal interest to us: (reject H_1 for H_2, 0, 0.1) and (reject H_2 for H_1, 0, 0.1). These are equally interesting and informative, and either one is preferable to the outcome (reject H_1 for H_2, 0.05, 0.05). But the latter can be made up of a mixture of equal parts of the first two tests – for example, we might toss a coin and use the result of that toss to decide which of the two tests to perform. This can be done in two ways. It can be done in such a way that I (the experimenter) know which test I am using; or it can be done in such a way that I remain ignorant of which test I am using. In the former case, the decision function M consisting of applying one or the other of the first two decision functions will be just as valuable as either of those decision functions – because at the time of the experiment I will in fact be using just one of those two decision functions, either one of which is satisfactory to me, and I shall furthermore know which one I am using. This form of rule M is surely just as valuable as either of the first two rules. In the latter case I will be informed only to the effect that 'Rule M has led to the rejection of H_1 for H_2, 0.05, 0.05', and I will not know in fact which decision rule has been applied in the particular case at hand. This would correspond to the situation in which the experimenter is indifferent between (reject H_1 for H_2, 0, 0.1) and (reject H_2 for H_1, 0.1, 0) but reasonably prefers either to (reject H_1 for H_2, 0.05, 0.05). But it is difficult to see what possible scientific point there could be to setting things up this way.

Of course we can imagine a situation in which we have three alternatives in designing an experiment, corresponding to the three tests mentioned by Birnbaum. Suppose that we are interested in the genotype of a single specimen insect of a species that breeds but once. We can

obtain evidence regarding the genotype by breeding the specimen with one of a number of (different) pure strains of the same species, and observing the phenotypes of the resulting offspring. Obviously, since the specimen can be bred but once, we can only do one of these experiments. Let h_1 and h_2 be hypotheses concerning the two aspects of the genotype we are examining. If we breed our specimen S to a purebred specimen of strain A, then if h_1 is true, we will never get offspring of phenotype T_1; if h_2 is true, the probability is 0.9 that some of the offspring will be of type T_1. If we breed S to a purebred specimen of strain B, then if h_2 is true, we will never get offspring of type T_2; if h_1 is true, the probability is 0.9 that some of the offspring will be of type T_2. We can obtain the test characterized by (0.05, 0.05) in three ways: As described before, we can flip a coin to determine which experiment to perform; assuming that strains A and B are phenotypically indistinguishable, we can put S in a black bottle with a specimen of each and let nature take its course; or we can employ a third purebred strain C, with the property that if h_1 is true the offspring will all be of phenotype T_3 with probability 0.05, and if h_2 is true the offspring will all be of phenotype T_3 with probability 0.95. Intuitively, either of the first two experiments with strain A or strain B seems preferable scientifically to any of the tests characterized by (0.05, 0.05). It is possible to imagine circumstances under which the experiment with strain C (or the three-bugs-in-a-bottle experiment) would be preferable. Either possibility entails the denial of the linearity required by the epistemic utility model.

This does not, of course, cast doubt on the cogency of Birnbaum's counterexample to the use of expected epistemic utilities alone to characterize statistical evidence, since there are other ways in which a test characterized by the pair of numbers (0.05, 0.05) can arise than as a mixture of equal parts of tests characterized by (0, 0.1) and (0.1, 0). It does provide further evidence, if any were needed, that the evidential interpretation of statistical evidence must depend on more than the pair of numbers, and, indeed, more than a specification of the pair of hypotheses together with their respective error probabilities. The value and significance of statistical evidence depend on the indefinitely large part of the body of knowledge that is available *after* the test has been performed.

4. That the notion of expected epistemic utility turns out to be a snare and

a delusion, at least in the context of choosing between two hypotheses, does not mean that we cannot consider some mistakes more serious than others, some hypotheses more valuable than others. Suppose, for example, that we are considering the mean of a random quantity with known variance. We draw a sample from the population, compute the sample mean, and, provided that the appropriate randomness conditions are met, are left with a posterior distribution for the unknown mean. From that posterior distribution, given a level of practical certainty, we can compute an infinite number of hypotheses, each of which is 'practically certain'. There are as many as there are ways of choosing Borel sets such that the integral of the posterior distribution over that Borel set yields a number equal to what we have taken to be practical certainty. But there is one that is of more interest than any other – the one which Neyman, for example, focused his attention on. This is the *shortest* one. This is common sense. This is a building block of the whole theory of shortest confidence intervals. There is an obvious sense in which this hypothesis, of all those that are equally probable, gives us the most information.

Why not make this a universal criterion, and take the length of the interval in which a hypothesis asserts a parameter to lie to be an inverse measure of the epistemic utility of the hypothesis? For one thing, there is no obvious way to apply that to the case of testing one hypothesis against another. For another, we have tacitly assumed, in the illustrative example, that the hypothesis about the mean has no logical bearing on any other hypothesis – on anything else in our body of knowledge. But this is surely not always the case. It is perfectly easy to imagine, for example, that if the mean length of the dorsal fin on a certain fish is greater than a certain length, then a whole evolutionary hypothesis must be altered, having indirect effects in turn on a whole group of other statistical hypotheses. And Birnbaum's examples illustrate the futility of attempting to measure the epistemic utility of a hypothesis considered by itself.

Thus, although we can find intuitive measures of value in particular cases, the search for a general measure of epistemic utility which will allow the reduction of the evidential uses of statistics to the behavioral uses which have been so thoroughly explored seems doomed to failure. What we need is a relatively new body of theory concerning the evidential impact of statistical data on statistical hypotheses. Furthermore, in order to apply such a theory to instances of the sort Birnbaum cites – genetic

investigations – we require in addition an understanding of the grounds on which scientific theories and hypotheses come to be accepted. Neither of these things can be developed in short order, but until they are developed the sciences that depend heavily on the analysis of statistical data must not only remain on shaky foundations, but must be shot through – as they are in fact shot through – with elements about which there is unresolved and apparently unresolvable disagreement among the experts involved. It is not merely that it would be logically or philosophically satisfying to have a clear and acceptable account of the relation between statistical hypotheses and statistical data – it is also that the empirical content of a number of bodies of scientific knowledge will come to be changed in the light of such an account.

University of Rochester