Error Probabilities in Error

Author(s): Colin Howson

Source: *Philosophy of Science*, Dec., 1997, Vol. 64, Supplement. Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers (Dec., 1997), pp. S185-S194

Published by: The University of Chicago Press on behalf of the Philosophy of Science Association

Stable URL: https://www.jstor.org/stable/188402

Error Probabilities in Error

Colin Howson[†‡]
London School of Economics

The Bayesian theory is outlined and its status as a logic defended. In this it is contrasted with the development and extension of Neyman-Pearson methodology by Mayo in her recently published book (1996). It is shown by means of a simple counterexample that the rule of inference advocated by Mayo is actually unsound. An explanation of why error-probablities lead us to believe that they supply a sound rule is offered, followed by a discussion of two apparently powerful objections to the Bayesian theory, one concerning old evidence and the other optional stopping.

"Bayesian Statistical Methods. A Natural Way to Assess Clinical Evidence." Title of lead editorial, *British Medical Journal* (BMJ 1996)

1. Introduction. People have been practicing inductive inferences for a long time, and have evolved lots of informal rules for designing informative experiments. An articulated and sound logic underwriting and explaining these rules has been longer in coming. The eighteenth century saw in the probability calculus such a logic, but because it appeared to be a profligate producer of inconsistency it became discredited, and was superseded by alternative methodologies in this century. Now it has made a comeback, under the name of the Bayesian theory, while those alternatives themselves have come under heavy attack.

The Bayesian theory can justify its claim to be a genuine logic in having something analogous to a completeness and soundness theorem: *the probability axioms can be shown to be the complete syntax of consistent probability assignments.* Consistency (also known as coherence) means that your evaluations of fair odds are consistent in the

†Department of Philosophy, London School of Economics, Houghton Street, London WC2A 2AE, England.

‡Thanks are due to Lawrence Jackson, Milo Schield, and Peter Urbach for their help, and to the British Academy for financial assistance.

Philosophy of Science, 64 (Proceedings) pp. S185–S194. 0031-8248/97/64supp-0018$0.00

sense that they do not depend on the form in which the relevant gambles are presented, and hence are invulnerable to a Dutch Book; consistency is thus an extensional semantic criterion, like truth (Howson and Urbach 1993, Ch. 5).

This account is usually called 'subjective', because it offers no endogenous evaluation of the values of the probabilities (or the corresponding odds), save on the extremes of necessarily true and false propositions. The terminology is question-begging. Nobody would say that deductive logic is subjective just because it does not tell you how to evaluate truth-values on propositions other than necessarily true and false ones. It is true that we believe (or some of us believe) that there is an autonomous realm of facts which make the sentences in deductive inferences objectively true or false. On the other hand, there are bodies of information available to the agent relative to which (s)he makes the judgment of what odds are justified in an uncertain proposition. Are these evaluations subjective? There is no general theory of how to perform them—Carnap made that endeavor unfashionable. But again, there is no general theory of how to arrive at truth-values: the best we can do is guess, subject to the constraints of a suitable logic for evaluating such guesses, or an inductive logic as it is usually styled.

Now the boot can be put on the other foot, for it turns out that the logic of coherence is itself an inductive logic. Various consequences of the probability axioms tell us how we should assess the truth-values of hypotheses conditional on acquiring evidence. The most celebrated of these is Bayes's Theorem, together with its variants. One of these is especially germane to the present discussion. Let h be a hypothesis and e evidence; then

$$P(h|e) = \frac{P(h)}{P(h) + \dfrac{\Sigma P(e|h_i)P(h_i)}{P(e|h)}}$$

where $\{h,h_i: i = 1, \ldots\}$ is the partition of alternative explanations of e that we regard as exhaustive, at any rate for now. We can see that for given P(h), P(e|h), (1) will be maximized by an outcome e such that $P(e|h_i)$ is small wherever $P(h_i)$ is not negligible. This tells us that as far as determining the likely truth of h is concerned, the most informative experimental designs are those which will effectively rule out the explanation of a positive outcome by any plausible alternative to h.

That is as much of the Bayesian theory that concerns us for the present discussion. The salient points are that it is an inductive logic, in that it provides information about how appropriate experiments can provide evidence in favor of (and also against) specified hypotheses,

and it that it is demonstrably sound, according to its semantic criterion of coherence. I shall contrast it in this respect with the recent adaptation, due to Mayo (1996), of the Neyman-Pearson theory beyond its original domain of purely statistical testing. In this, called by Mayo the 'error-based' approach, the following inferential rule, which I shall henceforth refer to as (*), is proposed: '*Evidence e should be taken as good grounds for H to the extent that H has passed a severe test with e*'. (Mayo 1996, 177, my italics; Mayo uses the capital H, and I shall follow her while discussing her account). She also says that such evidence 'indicates the correctness of hypothesis H' (p. 64).

What does passing a severe test with e mean? The answer requires a preliminary notion of an ordering of outcomes in terms of how well they 'accord with' or 'fit' H. This notion is never defined explicitly but Mayo proposes as a 'minimal' condition that e does not fit H if e is improbable under H. There are problems with this: 12 heads in 26 tosses of a coin intuitively accords with the hypothesis that the tosses are independent with constant probability of 1/2 of yielding a head, yet it is highly improbable under that hypothesis. Perhaps relative probability (or density where appropriate) rather than just probability would be a better way of characterizing fit, but I shall not dwell on this, because it will not affect the subsequent discussion. At any rate, a passing outcome is one that fits or accords with H (p. 178).

"Passing a severe test with e" is now defined to mean that "There is a very low probability that test procedure T would yield so good a fit [as e], if H is false" (p. 180). In other words, the probability of an outcome as good as e or better in the fitness ordering is very improbable if H is false. If we let A_e be the set of outcomes fitting H at least as well as e, the conditions for H to pass a severe test with e can be stated concisely as

(i) e fits H; and
(ii) $p_{\neg H}(A_e)$ is very small.

I have used a lower case p here to distinguish the sorts of probabilities involved in this account of severe tests with Bayesian probabilities. The former are the sort of probabilities that Carnap distinguished as probabilities$_2$, and Mayo calls chances, i.e., objective statistical probabilities, while the latter are epistemic. However, this restriction to chances requires immediate amendment of (ii), since talk of the chance assigned A by $\neg$H is meaningless where the background information does not determine a unique chance distribution when H is false. The amended (ii) for such hypotheses demands instead a small chance $p_{H'}(A)$ for each alternative hypothesis H′ in whatever set of alternatives to H consistent with that information (those familiar with Neyman-

Pearson theory will recognize the criterion of selecting uniformly most powerful tests where they exist). It can plausibly be objected that since this logic does not tell us what may and may not be used as such background information, the charge of arbitrariness so often leveled at the Bayesian theory for countenancing exogenously determined prior probability distributions could with equal justice be turned on advocates of the error-probability scheme. Never mind. There are other features of the rule (*) which I want to highlight.

First, it raises questions which it does not answer. Exactly how small, for example, is a very small probability, given $\neg$H, of getting an outcome according with H at least as well as e? And why do we have to consider the probabilities of outcomes other than those actually observed? These questions would presumably be answered in an adequate argument for (*) itself. But—and contrast this with the Bayesian theory—though there are various claims made that it is by severe testing that we eliminate error, there is nothing resembling a proof that (*) is a sound rule. Nor could there be, since, as I shall show in the next section, *(*) is demonstrably unsound.*

2. An Unsound Rule. The usual way of demonstrating unsoundness is by means of a counterexample, and I shall now describe one. Though the precise formulation is due to Korb (1991), the general idea is familiar in the literature (Rosenkrantz 1977, 206). Consider a test for the presence of a particular disease. Let H be the hypothesis 'the disease is present in the (randomly-chosen) test-subject'. To foreclose questions about the precise constitution of the set A_e above, we shall assume that the test delivers only two outcomes 'positive' and 'negative' (where 'negative' means 'the disease is absent' and 'positive' means that it is present). Thus A_{positive} is simply the singleton {positive}. Suppose that adequate statistics exist to determine p_H(positive) and $p_{\neg H}$(negative). Indeed, suppose that p_H(positive) is large, say 1. The test, in clinicians' argot, has 0% false negatives; in Mayo's terminology, 'positive' fits H maximally well. So H passes the test with 'positive'. Suppose also that the test has excellent false-positive rates; i.e., $p_{\neg H}$(positive) is very small, say 5%. Thus H passes a severe test with 'positive', and hence by (*) 'positive' indicates the correctness of H. But now also suppose that the disease is known to have an extremely small incidence, say 1 in 1000, i.e., .001. In this example we can faithfully represent all the subscripted probabilities as appropriate conditional probabilities, and we can use the probability calculus to tell us what the chance is of H being true given a positive outcome. By Bayes's Theorem we infer that p(H|positive) = .0196. The criteria for an extremely stringent test are satisfied, yet far from indicating the correctness of H, the data give the

chance that a subject who tests positive has the disease as very small indeed, under 2%!

The counterexample shows clearly that despite the test's being as severe as you like, it is a mistake to suppose that the very (small) chance of a test's passing a hypothesis h when h is false is by itself any indicator of the correctness of h if h passes the test (we can clearly get the same sort of result however small p('positive'|¬h) is, as long as it is positive, by correspondingly adjusting the prior for h). Indeed, if you infer from the test's positive diagnosis to the presence of the disease you will be wrong nearly all the time.

3. What's Gone Wrong? As numerous pieces of research have shown, people find it difficult to interpret probabilities, and the main reason is that ordinary language lacks the requisite scope operators, particularly where conditional probabilities are concerned, and consequently equivocates. The error characteristics of the test above sound impressive because of the informal way a type II error is standardly described, as a small chance of the test passing a false hypothesis, which it is equally easy to read as a small chance of a false hypothesis passing the test. Indeed, the two descriptions are virtually equivalent in ordinary discourse. Thus the informal reporting of the type II characteristics of the test in the context of the present example could easily be taken to mean either (i) a small chance of the diagnosis being positive if the disease is absent (i.e., the traditional type II error); or (ii) a small chance of the diagnosis being positive *conditional* on the disease being absent; or (iii) a small chance of the disease being absent conditional on the diagnosis being positive. *But these are all quite different probabilities.* In order to 'indicate the correctness of h', in Mayo's terminology, the test should have the property (iii), whereas in fact it has property (i) (she writes it as (ii)), *not* property (iii). In fact, as we have seen, a small type II error as in (i) puts almost no limit on how large the chance in (iii) can be, and therefore how *large* what we might call the 'real' type II error can be (Schield 1996 contains an excellent discussion along similar lines). Error statisticians, aided and abetted by the equivocal nature of informal language, have for decades given us something quite different from what we want, which is a way of computing the degree of confidence we should invest in hypotheses given empirical evidence. The error probabilities characterizing the test do not do this, as we see. Only recently has common sense returned, common sense reduced to a calculus, to use Laplace's apt words, and now known as the Bayesian theory. Indeed, since Bayes's Theorem was used to compute the chance of H given the outcome 'positive', the Bayesian theory thereby accu-

mulates weight behind its claim to be regarded as the logic of inductive support.

It might be objected that the hypothesis in the example is a random variable, whereas a hypothesis of the sort philosophers of science usually discuss is not (Mayo several times claims that hypotheses are not random variables). The objection is both wrong and beside the point. It is wrong because there are models of Kolmogorov's axioms in which hypotheses are random variables (measurable functions): any hypothesis is a two-valued random variable in the appropriate space. The objection is beside the point since the error-probability conditions for a severe test of that particular hypothesis H are clearly satisfied; equally clearly, passing the test provides no indication of H's correctness. Indeed, the counterexample is so telling precisely because H is a random variable, possessing an empirically-based prior distribution.

The message is clear: rely on error probabilities only at your peril. As Mayo remarks: "To get at the underlying rationale of a methodological rule we ask: if experiments were allowed to violate freely the methodological requirement in question, would some . . . clearly unreliable argument [be] allowed?" (1996, 454). Yes, it would be if you make the sorts of inference Mayo wants to make *without regard to priors,* whether they are empirically backed or not. Even Popper, no friend of the Bayesian theory, implicitly conceded this, when he pointed out that with a prior of 0 for a universal hypothesis h it is incorrect to regard the passing of a test, any test, as indicating anything at all about the truth-value of h, or its performance in future tests (Popper 1972, 18–19).

4. Some Objections to Bayesianism. The explicit dependence on subjective prior probabilities is the usual starting point for attacks on the subjective Bayesian theory. As we have seen, however, they are indispensable if inductive inferences are not to fall prey to fallacy through relying solely on the deceptive testimony of likelihoods. Nor are prior distributions optional. There is powerful evidence, supplied in the standard derivations of a utility-plus-probability scale and elsewhere, that to have what are formally prior probabilities is a condition of consistency in your evaluation of uncertainty. The fact that these are usually implicit rather than explicit does not mean that they are not there.

A current focus of strong criticism of the Bayesian theory, regarded even by John Earman as a black eye for the theory (Earman 1992, 135) is the so-called 'old evidence' objection, first voiced by Clark Glymour (1980). He noted that if e is already known at the time h is formulated, then $P(e) = 1$; hence trivially $P(h|e) = P(h)$, and so e fails to support h (increase its probability). But now this looks very bad for the Bayes-

ian theory because it conflicts sharply with our intuitions in some very well-known examples drawn from the history of science, like the precession of Mercury's perihelion vis-à-vis Einstein's theory of General Relativity. I am going to claim that the problem as formulated by Glymour rests on an improper use of a formula, as improper as using the cancellation law of arithmetic when the term cancelled is 0.

That some part of the Bayesian theory is being misused is evident when it is seen that it follows from the analysis above that no evidence can be regarded by a Bayesian as supporting any hypothesis. For evidence is by hypothesis known and so $P(e) = 1$ always. Were it a valid use of the Bayesian formulas to plug $P(e) = 1$ into the calculation of posteriors, then the theory would not have lasted three minutes let alone three hundred years. Clearly, you are not entitled to substitute 1 for P(e) and P(e|h) in computing support by Bayes's theorem just because you know that e is true.

What are you supposed to do? A simple example will tell us. Suppose that e reports the result of a sequence of n tosses, with observed frequency of heads r. You want to compute the posterior distribution of p, the probability of a head, on e. The value of P(e) that is relevant in this computation is the integral of ${}^{n}C_{r}p^{r}(1-p)^{n-r}f(p)$, where f(p) is your prior density distribution. This is equivalent to saying that for the purposes of the computation you are acting on the information supplied by your conditional probabilities and priors on the assumption that e is not yet known. In other words, you are assuming counterfactually that you do not yet know e.

What is true in the way support is computed in this example holds generally. Of the much written about 'old evidence', virtually all of it claims that there is no way of evaluating the counterfactual probability of e. Mayo even calls the idea 'silly' (1996, 334). But it is not silly at all: we have just seen how it can be done by appeal to other of your beliefs: P(e) is equal to the total probability $\Sigma P(e|h_i)P(h_i)$, where $\{h_i\}$ is the family of those alternatives you propose as alternative explanations. It might be objected that if e has been known a long time all the priors will be contaminated by this knowledge. This is a reasonable point. One way of decontaminating them is to assume for the purposes of this inference a uniform distribution, tantamount to letting the data decide between the competing alternatives. Or there may be other considerations unconnected with e you might wish to introduce, which weigh strongly with you, like symmetry considerations, or simplicity, etc. But once you have settled on your prior distribution the problem, at least theoretically, is solved: you will have found your counterfactual probability of e, which in general will be less than 1.

Glymour does consider this way of doing things in the context of

the Mercury perihelion example, but denies its validity because the measured perihelion shift, changing as it did over the years, was insecure, and therefore P(e) was never equal to 1 (Glymour 1980, 88). This seems to be beside the point, which is how, when it is equal to 1, a Bayesian can compute a value for P(h|e) different from P(h). So let us suppose that P(e) is 1 in this case, and see how the computation of a 'counterfactual' P(e) might go, bearing in mind the considerations above. There are two principal alternative explanations of the observations, Newton's theory (h) and Einstein's theory (h′), and some background information at the time about relevant parameters like the distribution of matter in the solar system. O.K. $P(e) = P(e|h)P(h) + P(e|h')P(h') + k$ where k is some smallish positive number (h and h′ might after all both be false). P(e|h) is close to 0 and P(e|h′) rather more substantial. Assume now that P(h) and P(h′) are initially both equal to just under 1/2 before the tribunal represented by e. We therefore have P(e) as approximately equal to $1/2P(e|h')$.

Mayo, who says she is echoing Earman, objects that attempting to 'subtract out' current knowledge of e in the Mercury case founders on the fact that Einstein constructed his theory using e as a constraint. What founders on this observation, however, is not the procedure I have just described, but the idea that it is the surprise-value of seeing that e is deducible from h that generates support for h (this is a view considered sympathetically by Garber (1983) and others). The view I have given certainly does not founder on it. Why should the fact that I have constructed h with e in mind preclude an assessment of how well e is explained by h in comparison with alternative explanations?

The final objection I shall consider to the Bayesian theory, and with which Mayo herself makes considerably play, is focused on optional stopping (i.e., the termination of data collection once some desired characteristic of it has been achieved). For example, it is a consequence of the law of the iterated logarithm that with probability one continued repetition will eventually yield an overall outcome an arbitrary number of standard deviations from the mean determined by the hypothesis, call it h, that the mean of a normal distribution with known variance takes a particular value. If there is a uniform, and hence improper, prior distribution over the mean it is well-known that the posterior distribution is normal (and so proper) with the same variance. This has the consequence, pointed out by Armitage (Savage 1962, 72) and quoted by Mayo (1996, 342–343), that if h is true then continued repetitions will with probability one generate an overall outcome giving h an arbitrarily small posterior probability. Similarly, if h is false, the strong law of large numbers tells us that with probability one we shall obtain a result giving h an arbitrarily small posterior probability. Yet if the experiment is delib-

erately planned to proceed until such an outcome is obtained, and then stop, our intuitions tell us that the posterior probability arises from data highly biased against h. Thus the Bayesian theory seems capable of producing to order conflicts with obviously correct intuitions. Mayo: "This is Armitage's argument. No satisfactory answer has been forthcoming, nor is there one. Armitage is right" (1996, 354).

One idea that might occur is to accept the premise that the information that the test continues until the requisite value of the posterior probability is obtained is relevant information, and condition on it. Such a strategy is in effect recommended by Rosenkrantz (1977, 199). However, the idea does not work in this case, for it amounts to conditioning on the set of outcomes which will generate that posterior probability; but this is a set of probability one, and so it makes no difference whether it is conditioned on or not (for why Rosenkrantz's strategy will not work in a more general setting see Seidenfeld 1979, n. 4).

There is nevertheless more than one way of answering Armitage. Kadane, Schervish, and Seidenfeld (1996) point out that the certainty of a misleading result like this depends on the use of improper priors. They give an elementary argument from expectations to show that with countably additive, and hence proper, probabilities the agent's probability function will never generate such 'foregone conclusions' (i.e., with probability one; 1996, S283).

However, it is not difficult to see that the certainty of obtaining misleading results even with the improper prior is not at all the methodological disaster it sounds. The fact is that you still can't actually plan in any meaningful sense to get them. You may know with certainty that you will, but you don't know when since there is no upper bound on the length of time you might have to wait (indeed, it might even be infinitely long). There is an analogy with number theory. You know that after a finite number k of computations you will get a 'yes' answer from your personal Turing machine just in case n is in a given recursively enumerable set Q of integers, but if Q is not recursive you cannot compute a bound on k. Armitage's mathematics is correct, but all it says is that misleading results will occur although you can not predict when. This fact is neither surprising nor does it provide an algorithm for being able to produce them (cf. Good 1983, 135, final paragraph).

REFERENCES

BMJ (1996), *British Medical Journal* 313: 569.

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Garber, D. (1983), "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory", in J. Earman (ed.), *Testing Scientific Theories.* Minneapolis: University of Minnesota Press, pp. 99–131.

Glymour, C. (1980), *Theory and Evidence*. Princeton: Princeton University Press.
Good, I.J. (1983), *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press.
Howson, C. and P. Urbach (1993), *Scientific Reasoning: The Bayesian Approach*. 2nd ed. Chicago: Open Court.
Kadane, J.B., M. J. Schervish, and T. Seidenfeld (1996), "When Several Bayesians Agree that There Will Be No Reasoning to a Foregone Conclusion", *Philosophy of Science* 63 (Proceedings): S281-S289.
Korb, K. (1991), "Explaining Science", *British Journal for the Philosophy of Science* 42: 239–253.
Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Popper, K.R. (1972), *Objective Knowledge*. Oxford: Clarendon Press.
Rosenkrantz, R. (1977), *Inference, Method and Decision: Toward A Bayesian Philosophy of Science*. Dordrecht: Reidel.
Savage, L.J. (ed.), (1962), *The Foundations of Statistical Inference: A Discussion*. London: Methuen.
Schield, M. (1996), "Using Bayesian Inference in Classical Hypothesis Testing". *Proceedings of the Statistical Education Section*. Washington, D.C.: American Statistical Association.
Seidenfeld, T. (1979), "Why I am not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz", *Theory and Decision* 11: 413–440.