BAYESIAN STATISTICS 2, pp. 249-270
J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (Eds.)

Weight of Evidence: A Brief Survey

I. J. GOOD
Virginia Polytechnic Institute and State University

SUMMARY

A review is given of the concepts of Bayes factors and weights of evidence, including such aspects as terminology, uniqueness of the explicatum, history, how to make judgments, and the relationship to tail-area probabilites.

Keywords: BAYES FACTORS; BAYESIAN LIKELIHOOD; CORROBORATION; DECIBANS; TAIL-AREA PROBABILITIES; WEIGHT OF EVIDENCE.

1. INTRODUCTION

My purpose is to survey some of the work on weight of evidence because I think the topic is almost as important as that of probability itself. The survey will be far from complete.

"Evidence" and "information" are related but do not have identical meanings. You might be interested in *information* about Queen Anne but in *evidence* about whether she is dead. The expression "weight of evidence" is familiar in ordinary English and describes whether the evidence in favour or against some hypothesis is more or less strong. The Oxford English Dictionary quotes T.H. Huxley (1878, p. 100) as saying "The weight of evidence appears strongly in favour of the claims of Cavendish", but C.S. Peirce used the expression in the same year so I suspect it was familiar before that date. The expression "The weight of the evidence" is even the title of a mystery story by Stewart or Michael Innes (1944). Moreover, Themis, the Greek goddess of justice is usually represented as carrying a pair of scales, these being for weights of evidence on the two sides of an argument.

A jury might have to weigh the evidence for or against the guilt or innocence of an accused person to decide whether to recommend conviction; a detective has to weigh the evidence to decide whether to bring a case to law; and a doctor has to weigh the evidence when doing a differential diagnosis between two diseases for choosing an appropriate treatment. A statistician can be said to weigh evidence in discrimination problems, and also, if he is not a Neyman-Pearsonian, when he applies a significance test.

In all these examples it seems obvious that the weight of evidence ought to be expressible in terms of probabilities, although the appropriate action will usually or always depend on utilities as well. At least three books have both the words "probability" and "evidence" in their titles (Good, 1950; Ayer, 1972; Horwich, 1982), as well as Dempster's lecture at this conference, and this again shows the close relationship of the two topics.

I believe that the basic concepts of probability and of weight of evidence should be the same for all rational people and should not depend on whether you are a statistician. There

should be a unity of rational thought applying, for example, to statistics, science, law, and politics. This assumption will set the tone of my survey. No concept is fundamental if only statisticians use it.

Suppose now that we have some hypothesis or theory, H, and some event, evidence, or experimental result called E. For example, H, might be the hypothesis that an accused person is guilty of some crime, and then the negation of H, denoted by $\overline{H}$, means that he is innocent. Of course H and $\overline{H}$ will seldom be simple statistical hypotheses in this legal example. A Bayesian, of whatever kind, assumes that it is meaningful to talk about such probabilities as $P(E|H\&G)$, $P(E|G)$, and so on, where G denotes background information such as that it is bad for the health to have guns fired at one. To economize in notation I shall usually take G for granted and omit it from the notation, so that the various probabilities will be denoted by $P(E|H)$ etc. These probabilities might be logical probabilities, known as "credibilities", or they might be subjective (meaning personal) or multisubjective (multipersonal); and they might be partially ordered, that is interval-valued, with upper and lower values, or they might have sharp (numerical) values. Although I believe partially-ordered probabilities to be more fundamental than sharp values, as I have said in about fifty publications (for example, Good, 1950, 1962a), I shall base my discussion on sharp values for the sake of simplicity. The discussion could be generalized to partially-ordered probabilities and weights of evidence but only with some loss of clarity. Any such generalization, if it is valid, should reduce to the "sharp" case when the intervals are of zero width. I would just like to remind you that inequality judgments of weights of evidence can be combined with those of probabilities, odds and other ratios of probabilities, expected utilities and ratios of them, etc., for improving a body of beliefs (for example, Good, 1962a, p. 322). I have not yet understood Shafer's theory of evidence, which is based on Dempster's previous work on interval-valued probabilities. Aitchison (1968) seems to me to have refuted the approach that Dempster took at this time, at least in some circumstances. At any rate, as I said, the present paper is based on sharp probabilities.

Other possible names for weight of evidence are "degree of corroboration" and "degree of confirmation", but the latter name was spoilt by Carnap because he made the mistake of calling a credibility a "degree of confirmation" thus leading philosophers into a quagmire of confusion into which some of them have sunk out of sight. This disaster shows the danger of bad terminology. Moreover, the expression "weight of evidence" is more flexible than the other two expressions because it allows such natural expressions as "the weight of evidence is against H". It would be linguistically unnatural to say "the degree of corroboration (or confirmation) is against H".

2. DERIVATION OF THE EXPLICATUM

I intend presently (meaning "soon") to discuss the history of the quantitative explication of weight of evidence, but it will be convenient first to mention a method of deriving its so-called explicatum from compelling desiderata. Let $W(H:E)$ denote the weight of evidence in favour of H provided by E, where the colon is read "provided by". If there is some background information G that is given all along, then we can extend the notation to $W(H:E|G)$. I mentioned that partly to show that we cannot replace the colon by a vertical stroke.

It is natural to assume that $W(H:E)$ is some function of $P(E|H)$ and of $P(E|\overline{H})$, say $f[P(E|H), P(E|\overline{H})]$. I cannot see how anything can be relevant to the weight of evidence other than the probability of the evidence given guilt and the probability given innocence,

so the function f should be mathematically independent of $P(H)$, the initial probability of H. But $P(H|E)$, the final probability of H, should depend only on the weight of evidence and on the initial probability, say

$$P(H|E) = g[W(H:E), P(H)].$$

In other words we have the identity

$$P(H|E) = g\{f[P(E|H), P(E|\overline{H})], P(H)\}$$

On writing $P(H) = x$, $P(E) = y$, and $P(H|E) = z$ we have the identity:

$$z \equiv g\{f[\frac{zy}{x}, \frac{y(1-z)}{1-x}], x\}.$$

It can be deduced from this functional equation that f is a monotonic function of $P(E|H)/P(E|\overline{H})$ (Good, 1968, p. 141) and of course it should be an increasing rather than a decreasing function. If H and $\overline{H}$ are simple statistical hypotheses, and if E is one of the possible experimental outcomes occurring in the definitions of H and $\overline{H}$, then $P(E|H)/P(E|\overline{H})$ is a simple likelihood ratio, but this is a very special case. In general this ratio is regarded as meaningful only to a Bayesian. It could be called a ratio of *Bayesian likelihoods*.

We can think of weights of evidence like weights in the scales of the Goddess of Justice, positive weights in one scale and negative ones in the other. Therefore we would like weight of evidence to have the additive property

$$W[H:(E\&E')] = W(H:E) + W(H:E'|E), \tag{1}$$

provided that this does not force us to abandon the result already established that $W(H:E)$ is a function of $P(E|H)/P(E|\overline{H})$. We can in fact achieve (1), uniquely up to a constant factor, by taking

$$W(H:E) = \log \frac{P(E|H)}{P(E|\overline{H})}. \tag{2}$$

Now we can easily see, by four applications of the product axiom of the theory of probability, namely $P(A\&B) = P(A)P(B|A)$, that

$$\frac{P(E|H)}{P(E|\overline{H})} = \frac{O(H|E)}{O(H)} \tag{3}$$

where O denotes odds. The odds corresponding to a probability p are defined as $p/(1-p)$. Some numerical examples of the relationship between probability and odds are shown in Table 1. In ordinary betting terminology odds of 2 are called odds of 2 to 1 on, and odds of ½ are called odds of 2 to 1 against, while odds of 1 are called "evens".

TABLE 1. *Probability and Odds*

Probability	Odds
0	0
1/10	1/9
1/3	1/2
1/2	1
2/3	2
9/10	9
1	∞

The right side of equation (3) can be described in words as the ratio of the final odds of H to its initial odds, or the ratio of the posterior to the prior odds, or the factor by which the initial odds of H are multiplied to give the final odds. It is therefore natural to call it the *factor in favour of H provided by E* and this was the name given to it by A.M. Turing in a vital cryptanalytic application in WWII in 1941. He did not mention Bayes's theorem, with which it is of course closely related, because he always liked to work out everything for himself. When I said to him that the concept was essentially an application of Bayes's theorem he said "I suppose so". In current Bayesian literature it is usually called the *Bayes factor in favour of H provided by E*. Thus weight of evidence is equal to the logarithm of the Bayes factor. The Bayes factor and weight of evidence are Bayesian concepts because the probabilities $P(H)$, $P(H|E)$, $P(E|H)$, and $P(E|\overline{H})$ are all in general regarded as meaningless by anti-Bayesians.

The additive property (1) simplifies if E and F are both independent given H and given $\overline{H}$. This condition usually requires that both H and $\overline{H}$ should be simple statistical hypotheses, a point of which Herman Rubin reminded me privately after the lecture.

The formula (3) occurs in a paper by Wrinch and Jeffreys (1921, p. 387); and Jeffreys (1936), called weight of evidence "support", but in this book, Jeffreys (1939), he dropped this expression because he always assumed that $O(H) = 1$ so that there $W(H:E)$ reduced to the logarithm of the final odds. His motive in concentrating on this special case must have been to try to sell fixed rules of inference: his original aim was to arrive at rules defining impersonal credibilities though his judgments of these were inevitably personal to him (Good, 1962b, p. 556). Whether they will become highly multipersonal is an empirical matter.

The basic property of weight of evidence can be expressed in words thus:

"initial log-odds plus weight of evidence = final log-odds".

Incidentally Barnard (1949), who, independently of Turing and of Wald, invented sequential analysis, called log-odds "lods". Good (1950), following a suggestion of J.B.S. Haldane, called it "plausibility", but "log-odds" is short enough and is self-explanatory.

It is sometimes convenient to write $W(H/H':E)$, read "the weight of evidence in favour of H as compared with H', provided by E", as a shorthand notation for $W(H:E|H \vee H')$. Of course, if $H \vee H'$ is given then $\overline{H} = H'$.

The Fisher-Neyman factorability condition for sufficiency (Fisher, 1925, p. 713; Neyman, 1925) can be expressed in terms of weight of evidence. I'll express it in terms of a class **H** of hypotheses instead of in terms of parameters. Let $f(E)$ be some function of the evidence. If $W[H/H':f(E)] = W(H/H':E)$ for all pairs H,H' of hypotheses in the class **H**, then $f(E)$ is sufficient for the hypotheses. Here $f(E)$ need not be a scalar or vector; it might be a proposition. This is a Bayesian generalization of the concept of sufficiency because $W(H/H':E)$ is not always an acceptable concept for the non-Bayesian. It could be called Bayesian sufficiency or "efficaciousness" (Good, 1958). For legal purposes, $f(E)$ is a possible interpretation of what is meant by "the whole truth and nothing but the truth", when there are two or more hypotheses to be entertained. Of course approximate Bayesian sufficiency is all that can be demanded in a court of law. Anyone who swears to tell the whole truth has already committed perjury.

For applications of weight of evidence, apart from the many applications for Bayesian tests of standard statistical hypotheses, see Good (1983e, p. 161).

3. SOME HISTORY

In the draft of my talk I said that C.S. Peirce (1878), is an obscurely written paper, had failed to arrive at the correct definition of weight of evidence owing to a mistake (see

Good, 1981b). But I intend to amend this comment in the discussion, in my reply to Dr. Seidenfeld. Levi (1982) agrees that Peirce made a mistake although he thinks it was different from the one I thought he made. Levi points out that Peirce was anti-Bayesian and to some extent anticipated Neyman and Pearson.

The definition of weight of evidence as the logarithm of the Bayes factor was given independently of Good (1950) by Minsky and Selfridge (1961); and again independently Tribus (1969) used the term "evidence" for weight of evidence. Kemeny and Oppenheim (1952) used the expression "factual support for a hypothesis" (provided by evidence), and their desiderata led them to the formula

$$\frac{P(E|H)-P(E|\overline{H})}{P(E|H)+P(E|\overline{H})}$$

This is an increasing function of $W(H{:}E)$, namely $\sinh\{W(H{:}E)/2\}$.

The philosopher Karl Popper (1954) proposed desiderata for corroboration and he said (1959, p. 394) "I regard the doctrine that the *degree of corroboration or acceptibility cannot be a probability* as one of the most interesting findings of the philosophy of knowledge". This fundamental contribution to philosophy was taken for granted by a dozen British cryptanalysts eighteen years before Popper published his comment, the name used being "score" or "decibannage". Moreover we used the best explicatum, which is not mentioned by Popper although this explicatum satisfies his desiderata.

In 1970, at the Second World Congress of the Econometric Society, Harold Jeffreys said it had taken fifty years for his work with Dorothy Wrinch to be appreciated by the statistical community, and he predicted that it would be another fifty years before the philosophers were equally influenced (or words to that effect). Recently the slowness of professional philosophers to use the correct explicatum for degree of confimation (or corroboration), namely $W(H{:}E)$, has been exemplified by Horwich (1982, p. 53). He suggests two measures, $P(H|E) - P(H)$ and $P(H|E)/P(H)$, of which the latter had been used by J.L. Mackie (1963). Although both these explicata satisfy the additive property (1), neither is a function of $W(H{:}E)$. To see that $P(H|E)-P(H)$ is inappropriate as a measure of degree of corroboration consider (i) a shift of probability of H from 1/2 to 3/4, (ii) a shift from 3/4 to 1, and (iii) a shift from .9 to 1.15. In each case $P(H|E)-P(H) = 1/4$, but the degree of corroboration seems entirely different in the three cases, especially as case (iii) is impossible! A similar objection applies to Mackie's suggestion $P(H|E)/P(H)$ and to its logarithm. For further discussion of Horwich (1982) see the review by Good (1983b). That review contains some other applications of the concept of weight of evidence to philosophical problems such as that of induction.

The unit in terms of which weight of evidence is measured depends on the base of its logarithms. The original cryptanalytic application was an early example of sequential analysis. It was called Banburismus because it made use of stationery printed in the town of Banbury; so Turing proposed the name "ban" for the unit of weight of evidence when the base of the logarithm is 10. Another possible name, especially in a legal context, would be a "themis" partly because Themis was the goddess of justice, and partly because Themis is the name of the tenth satellite of Saturn. But "ban" is shorter, more convenient, and historically justified. Turing called one tenth of this a *deciban* by analogy with a *decibel* in acoustics, and we used the abbreviation *db*. Just as a decibel is about the smallest unit of difference of loudness that is perceptible to human hearing, the deciban is about the smallest unit of weight of evidence that is perceptible to human judgment. It corresponds to a Bayes factor of 5/4 because $\log_{10}5 = .70$ and $\log_{10}4 = .60$. A bit is 3.01 db.

When I arrived at Bletchley the work on Banburismus had been going for some weeks and the entries on the Banbury sheets were of the form 3.6 meaning 3.6 db. I proposed first that the decimal point should be dropped so that the entries would be in centibans, and better still that the unit could be changed to a half-deciban or *hdb* with little loss of accuracy. This very simple suggestion saved much writing and eyestrain, and probably decreased the number of arithmetical errors. It may have cut the time for Banburismus by a factor of 2 and time was of the essence. One of my colleagues, Alex Kendrick, suggested the name "bonnieban" for the *hdb*.

The concept of weight of evidence is formally related to the logit transformation, $x = \log [P/(1-P)]$, although here P is a *cdf*. I don't think it explains why the logit transformation is useful, but it might have suggested the transformation to Fisher because Jeffreys was one of his students. Jeffreys (1936) discussed the logarithm of the Bayes factor and Fisher & Yates (1938) suggested the logit transformation.

4. A SIMPLE EXAMPLE

As a simple example, suppose we are trying to discriminate between an unbiased die and a loaded one that gives a 6 one third of the time. Then each occurrence of a 6 provides a factor of $\frac{1/3}{1/6} = 2$, that is, 3 *db*, in favour of loadedness while each non-6 provides a factor of $\frac{2/3}{5/6} = \frac{4}{5}$, that is, 1 *db*, against loadedness. For example, if in twenty throws there are ten 6's and ten non-6's then the total weight of evidence in favour of loadedness is 20 *db*, or a Bayes factor of 100. The corresponding tail-area probability for getting ten or more 6's if the die is fair is about 1/1670. A Bayes factor is always smaller than the reciprocal of a tail-area probability (Good, 1950, p. 94), and in this example it is smaller by a factor of 16.7.

5. HOW TO MAKE JUDGMENTS

Even when H and $\overline{H}$ are simple statistical hypotheses, in which case a Bayes factor is equal to a likelihood ratio, the terminology of Bayes factors and weights of evidence has more intuitive appeal. This intuitive appeal persists in the general case when the weight of evidence is not the logarithm of a likelihood ratio. I conjecture that juries, detectives, doctors, and perhaps most educated citizens, will eventually express their judgments in these intuitive terms. In fact, in legal applications, it must be less difficult to judge $P(E|H)/P(E|\overline{H})$ or $O(H|E)/O(\overline{H})$, or its logarithm, than to judge $P(E|H)$ and $P(E|\overline{H})$ separately because these probabilities are usually exceedingly small, often less than 10^{-100}. Of course the official responsibility of juries is more to judge $P(H|E)$ if they think in terms of probabilities. They are supposed to exclude some kinds of evidence, such as previous convictions, but they probably do allow for these convictions when they know about them, judging by some experiences of Hugh Alexander (c. 1955) when he served on a British jury.

A problem that arises both in legal and medical applications is in deciding what is meant by the initial probability of H. For example, if the accused is regarded as a random person in the world, his initial probability of guilt in much smaller than if he is known to live in the town, or village, where the crime was committed. For this reason it might often be easier to judge $O(H|E)$ directly than to compute it as $O(H)F(H:E)$ where F denotes the Bayes factor. Perhaps the best judgmental technique is to split the evidence into pieces and to check your judgments for consistency. For example, you could make separate judgments of (i) $O(H|E\&E')$ and (ii) $O(H|E)F(H:E'|E)$, while realizing that these should be equal. Some people, after some training, might find it easier to work with the additive

logarithmic form, instead of or in addition to the multiplicative form, that is, to use log-odds and weights of evidence (or log-factors) instead of odds and Bayes factors. It is convenient that factors of 2, 4, 5, 8, 10 and 20 correspond closely to weights of evidence of 3, 6, 7, 9, 10 and 13 decibans respectively. Themis should be grateful to Zeus for giving us just ten fingers.

I believe that this device of splitting the evidence into two or more groups corresponds to a psychologically natural manner of evaluating evidence. First some evidence E makes you suspicious, and you estimate the odds of H as somewhere near evens; then some more or less independent evidence arrives, perhaps in the form of a new witness, and this peps up the odds by a factor that you can judge separately. (Similarly an antibayesian, unaware that he is really a Bayesian, will choose null hypotheses of non-negligible prior probabilities, and then test them). It might help the judgment to recognize consciously that the chronological order of hearing evidence is not entirely relevant, and to imagine that it had arrived in some other order. In legal applications, one example of a convenient piece of evidence, that can be mentally separated from the rest, is the discovery of a strong motivation for the crime. An alibi is another example, in the opposite direction.

6. EXPECTATIONS AND MOMENTS OF BAYES FACTORS AND OF WEIGHTS OF EVIDENCE. ENTROPY

In 1941, or perhaps in 1940, Turing discovered a few simple properties of Bayes factors and weights of evidence. One curious result, which was independently noticed by Abraham Wald, was, in Turing's words "The expected factor in favour of a wrong hypothesis is 1". This fact can be better understood from its very simple proof: Suppose the possible outcomes of an experiment are $E_1, E_2, E_3, \ldots$ and that the hypothesis H is true. If E_i is an observed outcome the factor against H is

$$F(\bar{H}:E_i) = \frac{P(E_i|\bar{H})}{P(E_i|H)}.$$

Its expectation given the *true* hypothesis H is

$$\mathrm{E}[F(\bar{H}:E_i)|H] = \sum_i \frac{P(E_i|\bar{H})}{P(E_i|H)} \cdot P(E_i|H)$$

$$= \sum_i P(E_i|\bar{H}) = 1. \tag{4}$$

This result seems surprising at first sight, and not just because of its simplicity. If $\bar{H}$ is false we expect the Bayes factor in its favour to be less than 1 in most experiments. The only way to get an expected value of 1 is if the distribution of the Bayes factor is skewed to the right, that is, when the factor against the truth exceeds 1 it can be large.

To exemplify (4), let's consider the example concerning a die that we considered before and suppose that the die is really a fair one. Then, on one throw of the die, there is a probability of 1/6 that the factor in favour of loadedness is $\frac{1/3}{1/6} = 2$ and a probability of 5/6 that the factor in favour of loadedness will be 4/5. Hence the expected factor in favour of loadedness when the die is unloaded is $1/6 \times 2 + 5/6 \times 4/5 = 1/3 + 2/3 = 1$. Thus Turing's theorem can be used as a check of the calculation of a Bayes factor.

It is disturbing that one can get a large factor against the truth. This point will emerge again later in this talk.

Let $f = F(H{:}E)$. Then the nth moment of f about the origin given H is equal to the $(n+1)$st moment of f given $\bar{H}$; that is,

$$\mathbf{E}(f^n|H) = \mathbf{E}(f^{n+1}|\bar{H}). \tag{5}$$

The case $n = 0$ is Turing's result, just discussed. It can be further proved that $E(f^\alpha|H)$ is an increasing function of α for $\alpha > 0$. Better results will be published elsewhere.

This follows from an algebraic inequality that might date back to Duhamel & Reynaud (1823, p. 155); see Hardy, Littlewood, and Pólya, (1934, p. 26). By letting $\alpha \to +0$ we find, as I shall show in a moment, that

$$\mathbf{E}[W(H{:}E)|H] \geq 0 \tag{6}$$

or in words, the expected weight of evidence in favour of the truth is non-negative, and vanishes only when $W(H{:}E) = 0$ for all E of positive probability. This is of special interest because weight of evidence is additive so its expected value is more meaningful than that of a Bayes factor. This inequality was pointed out to me by Turing in 1941, with a different proof. Regarded as a piece of algebra it is merely an elementary inequality. What makes it interesting is its interpretation in terms of human and therefore statistical inference. That is why I regard it as reasonable to attribute it to Turing although it was also applied to statistical mechanics by Gibbs (1902, p. 136).

The monotonic property of $\mathbf{E}(f^\alpha|H)$ can be written

$$\Sigma \frac{p_i^{\alpha+1}}{q_i^\alpha} \text{ increases with } \alpha \quad (\alpha \geq 0) \tag{7}$$

where $p_i = P(E_i|H)$, $q_i = P(E_i|\bar{H})$.
But the left side is 1 when $\alpha = 0$, so

$$\Sigma p_i \left(\frac{p_i}{q_i}\right)^\alpha \geq 1,$$

that is,

$$\Sigma p_i \exp(\alpha \log \frac{p_i}{q_i}) \geq 1.$$

Therefore

$$\Sigma p_i [1 + \alpha \log \frac{p_i}{q_i} + \ldots) \geq 1.$$

By taking α small we get

$$\Sigma p_i \log \frac{p_i}{q_i} \geq 0 \tag{8}$$

which states that $\mathbf{E}(\log f|H) \geq 0$. Thus (7) can be regarded as a generalization of (6) or (8). The fact that (8) is an algebraic theorem confirms that weight of evidence is correctly explicated, although I hope you are already convinced. See also Good (1983d).

One way to interpret (6) or (8) is that in expectation it pays to acquire new evidence, if arriving at the truth is your objective. An explicit proof in terms of decision theory, that it pays in your own expectation to acquire new information, without reference to weight of evidence, was given by Good (1967, 319-321); but see also Good (1974) (where it was shown that, in some one else's expectation, it does not necessarily pay you). The principle is related to what Carnap (1947) called "the principle of total evidence": Locks' recommendation to use all the available evidence when estimating a probability.

Turing's inequality can of course be interpreted in terms of discrimination between two multinomials, a familiar problem in cryptanalysis. If $p_1, p_2, \ldots, p_n$ are the category

probabilities under the true hypothesis H, and are $q_1, q_2, \ldots, q_n$ under hypothesis $\overline{H}$, then the expression (8) is equal to the expected weight of evidence "per letter" in favour of H.

Sometimes one of the hypotheses is that of equiprobability, say that $q_1 = q_2 = \ldots = q_n = 1/n$. Then the expected weight of evidence becomes $\Sigma p_i \log(np_i)$ and this is equal to $\log n + \Sigma p_i \log p_i$. The expression $-\Sigma p_i \log p_i$ is usually called "entropy" because it is a form that entropy often takes in statistical mechanics (Boltzmann, 1964, p. 50; Gibbs, 1902, p. 129). That is because it is convenient for some purposes to divide phase space into equal volumes, in virtue of Liouville's theorem. (In thermodynamics, which is explained by statistical mechanics, the entropy has a different definition). The entropy expression $-\Sigma p_i \log p_i$ occurs prominently in Shannon's theory of communication, but his coding theorems can be somewhat better expressed in terms of expected weight of evidence in my opinion (Good & Toulmin, 1968).

Apart from its central position in human inference, one reason that expected weight of evidence is more fundamental than entropy is that it is applicable to continuous variables without ambiguity. This fact is related to its "splitative" property in the discrete case. That is, $\Sigma p_i \log(p_i/q_i)$ is unchanged if one of the categories is split into two categories in a random manner such as by spinning a coin. Among its names apart from "expected weight of evidence" are "relative entropy", "cross-entropy", "dinegentropy", and "discrimination information". (Should this one be "discrimination evidence"?)

The symmetrized expression $\Sigma(p_i - q_i)\log(p_i/q_i)$, called "divergence" by Kullback (1959), had been used also by Boltzmann (1964, p. 54) in statistical mechanics, Jeffreys (1946) in his theory of invariant priors, and by me for discrimination problems in cryptanalysis during World War II.

For deciding in advance whether Banburismus was likely to be successful, Turing estimated the expected weight of evidence. While doing so he discovered another theorem which I shall now state.

Suppose that the weight of evidence in favour of H, when H is true, has a normal distribution with mean μ and variance σ^2, and suppose our unit is the "natural ban". Then $\sigma^2 = 2\mu$. In other words

$$\text{var}\{W(H:E)\,|\,H\} = 2\,\mathbf{E}\{W(H:E)\,|\,H\}. \tag{9}$$

Moreover

$$\mathbf{E}\{W(\overline{H}:E)\,|\,H\} = -\mathbf{E}\{W(H:E)\,|\,H\} = -\mu\,. \tag{10}$$

This result was later published by Peterson, Birdsall and Fox (1954) in connection with radar. The result is surprising so I shall give the proof.

Let x be an observed weight of evidence in natural bans. Since the weight of evidence tells us just as much as E does about the odd of H, we have

$$W(H:E) = W\{H:W(H:E)\} = W(H:x).$$

Therefore the ratio of the probability densities

$$\frac{P.D.(x\,|\,H)}{P.D.(x\,|\,\overline{H})} = e^x\,. \tag{11}$$

Assume that x (or rather the corresponding random variable) has the distribution $N(\mu,\sigma^2)$ so that the probability density of x, given H, is

$$\frac{1}{\sigma\sqrt{2\pi}} \exp -\left[\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Then by (11)

$$P.D.(x|\overline{H}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} - x\right].$$

Therefore

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} - x\right] dx = 1$$

and from this it follows that $\sigma^2 = 2\mu$ and also that the distribution of x given $\overline{H}$ is $N(-\mu,\sigma^2)$.

If we use decibans the formula $\sigma^2 = 2\mu$ becomes converted to $\sigma = \sqrt{\mu 20\log_{10}e} = \sqrt{8.686\mu} \cong 3\sqrt{\mu}$. Thus the standard deviation is much larger than one might have guessed, a fact that in the application to radar is disturbing. For example, if the expectation is 16 *db*, which corresponds to a Bayes factor of 40, there is a probability of 1/6 that the weight of evidence will exceed $16 + 3\sqrt{16} = 28$ *db*, corresponding to a factor of 600, and a probability of 1/6 that it will be less than 4 *db*, corresponding to a factor of only 2½. Also there is a chance of 1/740 that the Bayes factor against the truth will exceed 100.

For generalizations of this theorem of Turing's to the more realistic case where it is assumed that the weight of evidence is only approximately normally distributed near its mean see Good (1961), which dealt with false-alarm probabilities in signal detection. The results can then be even more disturbing than in the case of strict normality and I hope this fact is well known to the defence departments of all countries that are civilized enough to possess an atom bomb.

Good & Toulmin (1968, Appendix B) and Good (1983f) give other relationships between the moments and cumulants of weight of evidence for the general case. Such identities can be deduced from the elegant formal identity $\phi(t+i) = \overline{\phi}(t)$ where ϕ and $\overline{\phi}$ denote the characteristic functions of $W(H:E)$ given H and $\overline{H}$ respectively. For example, when the moments exist,

$$\overline{\mu}_s' = \sum_{\nu=0}^{\infty} (-1)^\nu \mu_{s+\nu}'/\nu! = e^{-\Sigma}\mu_s', \qquad \mu_s' = \sum_{\nu=0}^{\infty} \overline{\mu}_{s+\nu}'/\nu! = e^{\Sigma}\overline{\mu}_s',$$

where μ_s' and $\overline{\mu}_s'$ denote moments about 0, and where Σ, just here, denotes the suffix-raising operator. There are similar identities for the cumulants. The cases $s = 0$ and $s = 1$ are of special interest.

When Turing judged the value of Banburismus by estimating an expected weight of evidence he was in effect treating weight of evidence as if it were a utility. It may be regarded as a *quasi-utility*, that is, an additive substitute for utility expressed in terms of probabilities. If you recall Wald's theorem that a minimax procedure is one that can be regarded as using a least favourable prior, you are led to the idea of minimizing expected weight of evidence or maximizing entropy in the selection of a prior. (Compare Good, 1969; Bernardo, 1979). Although minimax procedures in statistical inference are controversial they have the advantage of having invariant properties. The idea of using maximum entropy for choosing a prior was suggested by Jaynes (1957), though without mentioning the minimax property. For a recent statement of my views on maximum entropy see Good (1983c).

In the design of an experiment the entire distribution of weight of evidence, and in particular its variance, is of interest, and not just its expectation. In this respect weight of evidence, like money, is not an exact substitute for utility.

Expected weight of evidence is basic to the non-Bayesian approach to significance testing of Kullback (1959).

For some relationships between expected weight of evidence and errors of the first and second kinds see Good (1980). Other properties of weight of evidence can be located through the indexes of Good (1983e).

7. TAIL-AREA PROBABILITIES

A Fisherian might try to interpret weight of evidence, in its ordinary English sense, in terms of tail-area probabilities in tests of significance. Suppose then that a client comes to a Fisherian with experimental results E and he wants to know how much evidence this provides against some null hypothesis H, or even whether H is supported if that is possible. The client does not want to reject H too readily for he considers it to be simpler than its rivals and so easier to work with. For example, if he did not have experimental results he would have "accepted" H in the sense of assuming that its observational implications were approximately correct. (Should this be the definition of a null hypothesis?) The situation occurs, for example, when other hypotheses involve additional parameters. This by the way explains why it is not always better to replace a significance test by an estimation procedure. This point was made, for example, by Arnold Zellner in discussion at the 21st SREB-NSF Meeting on Bayesian Inference in Econometrics in 1980 in response to someone who was trying to knock significance tests. For several of my own views concerning significance tests see Good (1981a).

Suppose you apply a significance test and get a tail-area probability or P-value of .0455. (It is irrelevant to most of my discussion whether this is a single or double-tail.) How should you report this to your client? The answer is not as simple as it seems. You might report this result to the client in one of the following ways, depending on your philosophy and on the client's philosophy, and on the practical background of the problem:

(i) "The hypothesis is 20 to 1", as in the lines from War of the Worlds: "The chances of anything coming from Mars are a million to one. But still they come!" I hope it's not a million to one *on*!

(ii) "The odds against the hypothesis are about 20 to 1" (a familiar fallacy perpetrated by reputable scientists).

(iii) "The probability of getting so extreme an outcome is .0455 if the null hypothesis is true", where the meaning of "more extreme" needs to be stated. It can't mean that the probability density is small because the density can be made arbitrarily small, even where the mode originally occurred, by applying a suitable transformation to the independent variable. (Compare the usual attack against "Bayes's postulate" of a uniform distribution!).

(iv) "The probability of getting so extreme a result is less than .05 if the null hypothesis is true."

(v) "Reject H."

(vi) "Reject H because $P < .05$."

(vii) "I wouldn't reject H (as a good approximation) because H is *a priori* so probable." For example, suppose a coin gave 61 heads and 39 tails, H being the hypothesis that the coin is fair. (Here the double-tail-area, allowing for a continuity correction, is .036.)

(viii) "I wouldn't reject H because the cost to you of assuming $\overline{H}$ is too great."

(ix) "The result is not decisive: collect more data if it is practicable."

(x) "You should have consulted me in advance so that we could have decided on a rejection procedure in the Neyman-Pearson fashion."

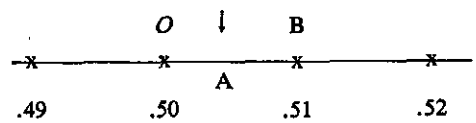
(xi) Ask the client "At what threshold P-value would you reject H?" Or arrive at a rejection level by discussion with the client.

(xii) None of the above.

I'm going to consider an example where «None of the above» is appropriate because the null hypothesis should be clearly accepted, not rejected. Let's imagine that the following game is being played at a gambling casino. An urn is known to contain 100 black and white balls. You pay an entrance fee, and the game consists in extracting one ball at a time. You win a dollar whenever a black ball is extracted. After each gamble the ball is returned to the urn and the urn is well shuffled, so the sampling is with replacement. Assume that each ball has probability 1/100 of being selected. Suppose that the game is played N times and that there are r successes and $N - r$ failures. We formulate the null hypothesis that there are 50 balls of each colour.

We are dealing with a binomial sample, and the standard deviation of r, given the null hypothesis, is $\sqrt{[N½(1-½)]} = ½\sqrt{N}$. For convenience assume that N is a perfect square and that $r = ½N + \sqrt{N}$. Thus the bulge is 2σ and the double tail-area probability $P = .0455$ so the result is «significant at the 5 % level». (I'm ignoring the continuity correction). I am now going to prove uncontroversially and without explicit Bayesianity that if N is large enough this outcome does not undermine the null hypothesis, in fact it supports it. This shows that it is incorrect to say that a null hypothesis can never be supported but can only be refuted, as one so often hears.

In this problem, the possible values of the binomial parameter p are 0, .01, .02,...,.99, 1.00, though the values 0 and 1 will have been ruled out if $r \neq 0$ or N.



In this diagram the possible values of p are marked with crosses. The observed fraction r/N of successes is marked by an arrow at the point A. The null hypothesis corresponds to the point O and the closest possible value for p, to the right of $p = ½$, is at B where $p = .51$.

The point A corresponds to a fraction $r/N = ½ + N^{-1/2}$. Thus, if N is large enough, the distance OA is much shorter than the distance AB. It is therefore obvious that if N is large enough our tail-area probability of .0455 supports the null hypothesis and the null hypothesis becomes more and more convincing as $N \to \infty$, corresponding to this fixed tail-area probability. A similar argument can be used even if the binomial parameter is continuous but it is not so clear-cut. It shows that a given P-value means less for large N. (Jeffreys, 1948, p. 222; 1961, p. 248; Hill, 1982; Good, 1983a). A possible palliative is to use *standardized tail-areas*. That is, if a small tail-area probability P occurs with sample size N we could say it is *equivalent to a tail-area probability of* $P\sqrt{100/N}$ *for a sample size of 100* if this is also small (Good, 1982b). The topic is closely related to the possibility of "sampling to a foregone conclusion" by using optional stopping when tail-area probabilities are used without any Bayesian underpinning. The earliest reference I know for this form of cheating is Greenwood (1938) and other references are given by Good (1982a).

Here is a Bayesian solution to the problem of the one hundred black and white balls in an urn. If there were only one rival H_p to the null hypothesis $H_{1/2}$, the Bayes factor against $H_{1/2}$ would be

$$F(H_p|H_{1/2}:r) = \frac{\binom{N}{r} p^r (1-p)^{N-r}}{\binom{N}{r} 2^{-N}} = (2p)^r (2-2p)^{N-r}$$

$$= (1+q)^r(1-q)^{N-r} \qquad (\text{where } q = 2p-1)$$
$$= \exp[r \log (1+q) + (N-r) \log (1-q)]$$
$$= \exp[r(q - \tfrac{1}{2}q^2 + \ldots) - (N-r)(q + \tfrac{1}{2}q^2 + \ldots)]$$
$$\approx \exp(2q\sqrt{N} - \tfrac{1}{2}q^2 N) \qquad (\text{when } r = \tfrac{1}{2}N + \sqrt{N}).$$

If we wanted to compute an exact Bayes factor against $H_{1/2}$ we would need to take a weighted average of the Bayes factors corresponding to each p (or q) the weights forming a prior distribution $P(H_p)$. But we don't need to do this in the present case because we obtain the maximum weight of evidence against $H_{1/2}$ by ignoring all values of q except $2 \times .51 - 1 = .02$. Thus the factor against $H_{1/2}$ is *at most* $\exp(\frac{\sqrt{N}}{25} - \frac{N}{5000})$, corresponding to a weight of evidence of $\frac{\sqrt{N}}{25} - \frac{N}{5000}$ natural bans. Here is a small table:

TABLE 2. *Evidence in favour of* $H_{1/2}$ if $P = .0455$

N	Weight of evidence *in favour* of $H_{1/2}$	Bayes factor *in favour* of $H_{1/2}$
40,000	≥ 0	≥ 1
90,000	≥ 6 nat. bans	≥ 400
1,000,000	≥ 160 nat. bans	$\geq 3 \times 10^{69}$

Thus in this example it is possible to get a lot of evidence in favour of the null hypothesis under circumstances where a dogmatic use of the 5 % rejection level would be ludicrous.

The primary lesson to be learnt from this example is that tail-area probabilities need to be used cautiously. If you use tail-area probabilities, perhaps you should always make an honest effort to judge whether your use of them is in violent conflict with your judgment of the Bayes factor or weight of evidence against the null hypothesis. In human thought, weight of evidence is a more fundamental concept than a tail-area probability. There was no Greek goddess who rejected hypotheses at the 5 % level with one tail in each scale!

Berkson (1942) criticised the view that a small P-value is evidence against a null hypothesis. He admits that he used to adopt the usual view but argues, without mentioning Bayes or Jeffreys, that (i) a small P-value is not evidence against the null hypothesis unless an alternative can be suggested that would make this low value more probable; (ii) values of P in the range (.3, .7) can support the null hypothesis for large samples. There is an error on page 333 where he says that "small P's are more or less independent, in the weight of the evidence they afford, of the numbers in the sample". (See also page 332). Otherwise Berkson's paper was largely Bayesian although he didn't notice it. Everybody is to some extent a Bayesian especially when using common sense.

ACKNOWLEDGEMENT

This work was supported in part by an N.I.H. Grant number GM18770.

REFERENCES

AITCHISON, J. (1968). Comment on a paper by A.P. Dempster, *J. Roy. Statist. Soc. Ser. B* **30**, 234-236.

ALEXANDER, C.H.O'D. (c. 1955). Oral communication.

AYER, A.J. (1972). *Probability and Evidence*. Columbia University Press.

BARNARD, G.A. (1949). Statistical inference, *J. Roy. Statist. Soc. B* **11**, 115-149 (with discussion)

BERKSON, J. (1942). Tests of significance considered as evidence. *J. Amer. Statist. Assoc.* **37**, 325-335.

BERNARDO, J. (1979). Expected information as expected utility, *Ann. Statist.* **7**, 686-690.

BOLTZMANN, L. (1964). *Lectures on Gas Theory*, Berkeley: University of California Press. Trans. by Stephen G. Brush from the German, *Gastheorie* of 1896-1898.

CARNAP, R. (1947). On the application of inductive logic, *Philosophy and Phenomenological Research* **8**, 133-148.

DUHAMEL, J.M.C. & REYNAUD, A.A.L. (1823). *Problèmes et développements sur diverses parties des mathématiques*. Paris.

FISHER, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society*, **22**, 700-725.

FISHER, R.A. & YATES, F. (1938). *Statistical Tables for Biological, Agricultural, and Medical Research*. Edinburgh: Oliver and Boyd.

GIBBS, J.W. (1902). *Elementary Principles in Statistical Mechanics*. London: Constable; Dover reprint, 1960.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Charles Griffin; New York: Hafners.

— (1958). Significance tests in parallel and in series, *J. Amer. Statist. Assoc.*, **53**, 799-813.

— (1961). Weight of evidence, causality, and false-alarm probabilities, *Information Theory, Fourth London Symposium* (1960). London: Butterworth.

— (1962a). Subjective probability as the measure of a non-measurable set, *Logic, Methodology, and Philosophy of Science* (Nagel, E., Suppes, P., Tarski, A., eds.) California: Stanford University Press. Reprinted in *Studies in Subjective Probability*, 2nd. edn. (H.E. Kyburg & H. E. Smokler, eds.; Huntington, N.Y.: R.E. Krieger), 133-146; and in Good (1983e), 73-82.

— (1962b). Review of Harold Jeffreys *Theory of Probability*, Third edn. London: Oxford University Press. *The Geophysical Journal of the Royal Astronomial Society*, **6**, 555-558. Also in *J. Roy. Statist. Soc. Ser. A*, **125**, 487-489.

— (1967). On the principle of total evidence, *British Journal for the Philosophy of Science* **17**, 319-321.

— (1968). Corroboration, explanation, evolving probability, simplicity, and a sharpened razor, *British Journal for the Philosophy of Science*, **19**, 123-143.

— (1969). What is the use of a distribution?, *Multivariate Analysis II* (ed. P.R. Krishnaiah); New York: Academic Press, 183-203.

— (1974). A little learning can be dangerous, *British Journal for the Philosophy of Science* **25**, 340-342.

— (1980). Another relationship between weight of evidence and errors of the first and second kinds, C67 in *Journal of Statist. Comput. & Simul.* **10**, 315-316.

— (1981a). Some logic and history of hypothesis testing, in *Philosophy in Economics* (ed. Joseph C. Pitt); Dordrecht: D. Reidel, 149-174. Also in Good (1983e), 127-148.

— (1981b). An error by Peirce concerning weight of evidence. *Journal of Statist. Comput. & Simul.* **13**, 155-157.

— (1982e). Comment on a paper by G. Shafer. *J. Amer. Statist. Assoc.* **77**, 342-344.

— (1982b). Standardized tail-area probabilities, **C140** in *Journal of Statist. Comput. & Simul.* **16**, 65-66.

— (1983a). The diminishing significance of a fixed *P*-value as the sample size increases: a discrete model, **C144** in *Journal of Statist. Comput. & Simul.* **16**, 312-314.

— (1983b). Review article of Paul Horwich, *Probability and Evidence*, Cambridge University Press, 1982; *British Journal for the Philosophy of Science*, **35**, 161-166.

— (1983c). Review of "The Maximum Entropy Formalism", Raphael D. Levine & Myron Tribus, eds. (1979). *J. Amer. Statist. Assoc.*, **78**, 987-989.

— (1983d). When are free observations of positive expected value?, **C161** in *Journal of Statist. Comput. & Simul.*, **17**, 313-315.

— (1983e). *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota Press.

— (1983f). Moments and cumulants of weights of evidence, **C162** in *J. Statist. Comput. & Simul.*, **17**, 315-319; **18**, 85.

GOOD, I.J. & TOULMIN, G.H. (1968). Coding theorems and weight of evidence, *J. Inst. Math. Applics.* **4**, 94-105.

GREENWOOD, J.A. (1938). An empirical investigation of some sampling problems, *Journal of Parapsychology* **2**, 222-230.

HARDY, G.H., LITTLEWOOD, J.L. & PÓLYA, G. (1934). *Inequalities*. Cambridge: University Press.

HILL, B.M. (1982). Comment on a paper by G. Shafer. *J. Amer. Statist. Assoc.* **77**, 344-347.

HORWICH, P. (1982). *Probability and Evidence*. Cambridge: University Press.

HUXLEY, T.H. (1878). *Physiography: an Introduction to the Study of Nature*. London: Macmillan; New York: D. Appleton. 2nd edn.

JAYNES, E.T. (1957). Information theory and statistical mechanics, *Physical Review* **106**, 620-630; **108**, 171-190.

JEFFREYS, H. (1936). Further significance test, *Proceedings of the Cambridge Philosophical Society*, **32**, 416-445.

JEFFREYS, H. (1939/1948/1961). *Theory of Probability*. Oxford: Clarendon Press.

— (1946). An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society, A.*, **186**, 453-461.

KEMENY, J.C. & OPPENHEIM, P. (1952). Degrees of factual support, *Philosophy of Science*, **19**, 307-324.

KULLBACK, S. (1959). *Information Theory and Statistics*. New York: Wiley.

LEVI, I. (1982). Private communication, December 9.

MACKIE, J.L. (1963). The paradox of confirmation. *British Journal for the Philosophy of Science* **13**, 265-277.

MINSKY, M. & SELFRIDGE, O.G. (1961). Learning in random nets, in *Information Theory: Fourth London Symposium* (Colin Cherry, ed.). London: Butterworhts, 335-347.

NEYMAN, J. (1925). Su un teorema concernente le cossidetti statistiche sufficienti, *Giorn. Inst. Ital. Attuari*, **6**, 320-334.

PEIRCE, C.S. (1878). The probability of induction, *Popular Science Monthly*, reprinted in *The World of Mathematics*, **2**, (ed. James R. Newman); New York: Simon and Schuster, 1956, 1341-1354.

PETERSON, W.W., BIRDSALL, T.G. & FOX, W.C. (1954). The theory of signal detectability, *Trans. Inst. Radio Engrs. PGIT-4*, 171-212.

POPPER, K. (1954). Degree of confirmation, *British J. Philosophy Sc.* **5**, 143-149.

— (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

STEWART, J.I.M. ("Michael Innes") (1944). *The Weight of the Evidence* London: Gollancz; also Harmondsworth, Middlesex, England: Penguin Books, 1961.

TRIBUS, M. (1969). *Rational Descriptions, Decisions and Design*. New York: Pergamon Press.

WRINCH, D. & JEFFREYS, H. (1921). On certain fundamental principles of scientific inquiry, *Philosophical Magazine, Series 6*, **42**, 369-390.

H. RUBIN (*Purdue University*)

Possibly some of my difficulties with philosophy are due to my allergy to horseradish (see Professor Seidenfeld's comments). However, some philosophy is necessary.

The consideration of situations in which the state of nature is highly restricted is necessary to clarify thinking. However, one must resist the temptation, made 99.99 % of the time by users of statistics, to believe the model. There is no conceivable way that I can state my prior or posterior probabilities in the last example in the paper; all that can be said is that it is reasonable (or unreasonable) to *act* as if the results come from Bernoulli trials with probability .5. I completely agree with Professor Good that a "significant difference" is *not* the proper criterion here. If there was a relative frequency of 50.1 % in 10^6 trials, on this evidence I personally would "accept" the hypothesis; with 10^{12} trials I would reject it; and with 10^9 trials I would think hard about the matter.

The model given is, in practice, *never* correct. Thus we can only use the evidence to decide which actions or statements to make. If a hypothesis is broad enough, it can be true; if it is too specific, it must be false, but it may still be appropriate and reasonable to act as if it is true.

T. SEIDENFELD (*Washington University, St. Louis*)

As a philosopher interested in "foundations", I take delight in the opportunity to comment on the papers of our distinguished speakers. Let me preface these remarks, more in the form of questions, with an admission of my perception of the role of philosophy in a session titled "Probability and Evidence". To paraphrase Larisa in Pasternak's *Dr. Zhivago* (chapter 13, §16), philosophy is like horseradish. It is good if taken in small amounts in combination with other things. But it is not good in large amounts by itself. The risk with philosophy, as with horseradish, is the temptation to use ever stronger concentrations to maintain the sensation of that first taste. Soon you are serving up pure horseradish!

Professor Good's savory recipe calls for a dash of philosophy in the form of an explication of "weight of evidence". Explication, you recall, is the business (made into an industry thanks to Carnap) of making clear and precise an inexact concept (the *explicandum*) taken from everyday language. The explicandum has all the obscurity typical of presystematic talk. In explication, the vague explicandum is replaced by an *explicatum* which, though similar to the original notion, must be exact, fruitful and simple. *Explication* is Carnap's (1950) explicatum for the explicandum "philosophical analysis".

Carnap begins his *Logical Foundations* in the hope of providing an explication of "probability". In 600 pages that follow, he struggles to defend the thesis of probability as a logical relation. In so far as Carnap's attempt at explication is not successful, I think it fair to say he does not meet the requirement of *usefulness*. Carnap's effort with logical probability fails to yield productive conceptual tools for reconstructing, e.g. statistical

inference. For one, he misses completely the important problem of the "reference class" for direct probability: how do we reconcile information from different statistical "populations" concerning some common "individual"?

I have a parallel concern with Good's explication. His account is exact and, no doubt, simple enough. But how does "weight of evidence" serve a useful purpose in solving problems of inference or decision? Let me argue, briefly, that two natural, candidate roles for an explication of *weight* are not fulfilled by Good's explicatum. Then it will be up to the author, himself, to point out what he intends for his creation.

J.M. Keynes, in chapter 6 of his *Treatise* (1921), raises the subject of weight of evidence along with the caveat that he remains uncertain how much importance to attach to the question. For Keynes, *weight of evidence* cannot be defined by probability as he sees weight monotonically increasing with increasing evidence. To use Keynes' metaphor (p. 77) *weight* measures the sum of favourable and unfavourable evidence whereas probability indicates the difference between these two. Keynes suspected that this hazy notion of *weight* plays a role in decisions separate from the role of probability. I do not think what Keynes had in mind requires a violation of expected utility theory. One interpretation of his query is to ask for a measure of weight of evidence that would help determine when a decision maker has adequate evidence for a (terminal) choice. That is, I propose we understand Keynes's problem with *weight* as his groping for a formulation of the stopping problem to which *weight* would offer the key to a solution.

In discussing the requirement of total evidence he writes,

> Bernoulli's second maxim, that we must take into account all the information we have, amounts to an injunction that we should be guided by the probability of that argument, amongst those of which we know the premisses, of which the evidential weight is greatest. But should not this be reenforced by a further maxim, that we ought to make the weight of our arguments as great as possible by getting all the information we can? It is difficult to see, however, to what point the strengthening of an argument's weight by increasing the evidence ought to be pushed. We may argue that, when our knowledge is slight but capable of increase, the course of action, which will, relative to such knowledge, probably produce the greatest amount of good, will often consist in the acquisition of more knowledge. But there clearly comes a point when it is no longer worth while to spend trouble, before acting, in the acquisition of further information, and there is no evident principle by which to determine *how far* we ought to carry our maxim of strengthening the weight of our argument. A little reflection will probably convince the reader that this is a very confusing problem. (pp. 76-77).

Some sixteen years ago, Good published a philosophical note (1967) in which he, like Keynes before him, connected the requirement of total evidence with the stopping problem. The upshot of that note is the result (also reported in Savage (1954, pp. 125-126)) that procrastination is best when observations are cost-free and not (almost surely) irrelevant. But, that finding as well as the general theory of optimal stopping is tangential to Good's concept of *weight*. Of course, with enough *weights* we recover the likelihood function. Hence, the *weights* are sufficient (though hardly a *reduction* of the data). Except in special cases, however, the stopping rule is not a function merely of the *weights*. Is there some reason to think Keynes was on the right track when he posited weight of evidence to solve optimal stopping? It seems to me current wisdom would label this a dead-end approach. Nor does Good's explicatum serve such a purpose.

A second role weight of evidence might conceivably play is in fixing belief. When is it reasonable to add a consistent belief on the basis of new evidence? An informal reply is: you are justified in coming to believe a proposition when the weight of the new evidence is

strong enough in its favor. Unfortunately, it seems Good's explicatum does nothing to defend this intuition.

The point is simple. Total evidence requires we respect equivalences implied by all we know. If $\mathbf{h}_1$ and $\mathbf{h}_2$ are equivalent given all our evidence, then whatever epistemic stance we take toward the one we take toward the other. To believe the one is to believe the other. But *weight of evidence* (here, of a kind with relevant measures of "support") does not conform to the needed invariance.

For example, let $X_i = 0,1$ $(i = 1,2)$ be two Bernoulli trials. Suppose the personal probability is symmetric and exchangeable and satisfies: $p(X_i = 0) = .5$ and $p(X_1 + X_2 = 0) = .05$. Hence, $p(X_1 + X_2 = 2) = .05$ and $p(X_1 + X_2 = 1) = .9$. Let $\mathbf{e}$ be the new evidence: $X_1 = 1$.

Let $\mathbf{h}_1$ be the hypothesis that $X_1 + X_2 = 2$ and $\mathbf{h}_2$ the hypothesis that $X_2 = 1$. Given $\mathbf{e}$, $\mathbf{h}_1$ and $\mathbf{h}_2$ are equivalent. But $\mathbf{e}$ has positive *weight* for $\mathbf{h}_1$ and negative *weight* for $\mathbf{h}_2$. If we use *weight* to account for our presystematic talk (weight measures reason for/against adding belief), then we have the incoherent conclusion that, given all we know, $\mathbf{e}$ is evidence for and against the same belief. It is an elementary and familiar exercise to show this phenomenon ubiquitous.

In a recent paper with D.Miller, Sir Karl Popper (1983) expresses concern over failure of "positive relevance" to respect such equivalences. Thus, I do not agree with Good (p. 8) when he speculates that *weight* satisfies Popper's desiderata for degree of corroboration or acceptability.

In short, my question to Professor Good is this one. What shall I do with *weight of evidence*?

REPLY TO THE DISCUSSION

Teddy Seidenfeld was kind enough to send a copy of his comments before the meeting. His main question was "What shall I do with weight of evidence". I think there must be some misunderstanding because my answer is so simple. My answer is that the weight of evidence provided by E should be added to the initial log-odds of the hypothesis to obtain the final log-odds. Or equivalently, the Bayes factor is multiplied by the initial odds to give the final odds. The final odds are then combined with utilities to make rational decisions. Weights of evidence and Bayes factors resemble likelihood, they have the merit of being independent of the initial probability of the hypothesis. Moreover the technical meaning of weight of evidence captures the ordinary linguistic meaning and that is my main thesis.

In the example, used by Seidenfeld to question the explication of weight of evidence, he had effectively the table of probabilities,

X_1 \ X_2	0	1	
0	.05	.45	.5
1	.45	.05	.5
	.5	.5	1

with $H_1: X_1 = X_2 = 1$ $\quad P(H_1) = .05, O(H_1) = 1/19$

$H_2: X_2 = 1$ $\quad P(H_2) = \frac{1}{2}, \ O(H_2) = 1$

$E: X_1 = 1$ $\quad$ (H_1 and H_2 are logically equivalent, given E).

	Bayes factor provided by E	*Initial odds*	*Final odds*
H_1	19/9	1/19	1/9
H_2	1/9	1	1/9

The fact that the final odds of H_1 and H_2 are equal, given E, is consistent with the fact that H_1 and H_2 are equivalent given E. The evidence E, that $X_1 = 1$, supported H by increasing its odds to 1/9, and undermined H_2 by decreasing its odds to 1/9. Before E was known, H_1 and H_2 were not equivalent and their initial odds were not equal. The ocurrence of E has simply changed the situation. Seidenfeld seems to have confused $W(H{:}E)$ with $W(H{:}E|E)$. The latter expression is equal to zero. That is, once E is *given* it supplies no further weight of evidence. To imagine that it does is like trying to double the true weight of evidence. The error is prevented by noticing the distinction between the vertical stroke which means "given" and the colon which means "provided by".

Popper made a different mistake when he apparently equated corroboration with acceptability, and when I said that his remark about corroboration had been previously taken for granted I should have made it clear that I was referring only to his statement that degree of corroboration cannot be a probability. My definition of weight of evidence does essentially satisfy all the desiderata for corroboration laid down by Popper in the Appendix dealing with the topic in his *Logic of Scientific Discovery* (Popper, 1959, pp. 400-401). The meaning of "essentially" here is spelt out in Good (1960, p. 321); for example, I replace Popper's bounds of ± 1 on degree of corroboration, by $\pm \infty$. Perhaps Popper has since shifted his position.

The alleged proof by Popper & Miller (1983) of the impossibility of inductive probability is unconvincing, and I have written a note to *Nature* arguing this (Good, 1983h).

A special case of weight of evidence was used by Peirce (1878) although he did not express it in Bayesian terms; in fact, as Isaac Levi has pointed out, Peirce anticipated the Neyman-Pearson theory to some extent. Incidentally, when a Neyman-Pearsonian asserts a hypothesis H, he unwittingly provides a Bayes factor of $(1-\alpha)/\beta$ in favour of H; and, when he rejects H, he similarly provides a Bayes factor of $(1-\beta)/\alpha$ against H. (See Good, 1983g; Wald, 1947, p. 41). These results are based on the assumption that we know the values of α and β, and we know what recommendation is made by the Neyman-Pearsonian, and nothing else. We can achieve this state of ignorance by employing a Statistician's Stooge who, by definition, is shot if he tells us more than we ask him to.

I have referred to practical applications in my paper, such as to sequential analysis, an example of which was Banburismus. The Bayes factor is also used throughout Harold Jeffreys's book on probability, though he nearly always assumes that the initial odds are 1. Every Bayesian test of a hypothesis can be regarded as an application of the concept of weight of evidence. Perhaps the most important applications, like those of probability, are the semiquantitative ones in the process of rational thinking as an intelligence amplifier.

Keynes's definition of weights of arguments, in which he puts all the weights in one scale, whether they are positive or negative, is like interpreting weight of evidence as the weight of the documents on which they are printed. I think, if not horseradish, it is at least a crummy concept in comparison with the explicatum of weight of evidence that I support. Keynes himself said of his discussion (1921, p. 71) "... after much consideration I remain uncertain as to how much importance to attach to it. The magnitude of the probability of an argument ... depends upon a *balance* between what may be termed the favourable and the unfavourable evidence...". In other words he clearly recognizes that Themis is right to use both scales. It is a standard English expression that the weight of evidence favours such

and such. Of course this refers to the *balance* of the evidence, *not* to the sum of all the pieces irrespective of their signs.

If you *must* have a quantitative interpretation of Keynes's "weight of arguments", just compute the weights of evidence in my sense for each "piece" of evidence and add their absolute values. This then is yet another application of my explicatum, to give a somewhat quantitative interpretation to the crummy one. But Keynes's discussion of this matter is purely qualitative.

Seidenfeld raises the question of whether weight of evidence can be used for deciding when to stop experimentation. My answer is that weight of evidence is only a quasiutility, as I stated in my paper. When you have a large enough weight of evidence, diminishing returns set in, where the meaning of "large enough" depends on the initial probability and on the utilities. Weight of evidence is a good quasiutility, and it is fine that the expected weight of evidence from an observation is nonnegative. But it cannot entirely replace expected utility as a stopping rule. When a judge's estimate of the odds that an accused person is guilty or innocent reaches a million-to-one on, the judge is apt to say "Finis" and bang down his gavel. This is because, in his implicit or explicit opinion, the expected gain in utility from seeking new evidence is not worth the expected time it would take to acquire.

I turn now to the public comments made by Herman Rubin. He stated that $P(E|H)$ is not well defined if H is a composite hypothesis. This is certainly true in non-Bayesian statistics but in "sharp" Bayesian statistics it is assumed to have a sharp value. For the sake of simplicity most of my exposition was based on the sharp Bayesian position. My aim was to discuss weight of evidence without going into the foundations of probability.

Dr. Rubin mentioned that, in my example of sampling with replacement from a bag of 100 black and white balls, if he obtained 50.1 % white drawings in a sample of a trillion, he would reject the model. So would I, but my example was based on a fixed P-value of 0.045. I deliberately selected not too small a P-value so that the model itself would not come under suspicion.

I agree further that precise models are seldom exact, but they are often useful on grounds of simplicity. Compare, for example, Good, 1950, p. 90; 1983e, p. 135.

REFERENCES IN THE DISCUSSION

CARNAP, R. (1950). *Logical Foundations of Probability*. Chicago: University Press.

GOOD, I.J. (1952). Rational decisions. *J. Roy. Statist. Soc. B.*, **14**, 107-114.

— (1955). Contribution to the discussion on the Symposium on Linear Programming. *J. Roy. Statist. Soc., B*, **17**, 194-196.

— (1960). Weight of evidence, corroboration, explanatory power, information, and the utility of experiments. *J. Roy. Statist. Soc., B*, **22**, 319-331; **30** (1968), 203.

— (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34**, 911-934.

— (1967). On the principle of total evidence. *British Journal for the Philosophy of Science*, **17**, 319-321.

— (1980). Some history of the hierarchical Bayesian methodology. *Bayesian Statistics* (Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M. eds.) Valencia: University Press, 489-510 & 512-519 (with discussion).

— (1983g). A correction concerning my interpretation of Peirce, and the Bayesian interpretation of Neyman-Pearson 'hypothesis determination'", C165, *J. Statist. Comput. & Simul.* **18**, 71-74.

— (1983h). The inevitability of probabilistic induction. *Nature* **310**, 434.

KEYNES, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

POPPER, K. and MILLER, D. (1983). A proof of the impossibility of inductive probability. *Nature* **302**, 687-688.

SAVAGE, L.J. (1954). *The Foundation of Statistics*. New York: Wiley.

SEIDENFELD, T. (1979). Why I am not an objective Bayesian: some reflections prompted by Rosenkrantz. *Theory and Decision 11*, 413-440.

WALD, A. (1947). *Sequential Analysis*. New York: Wiley.