Statistical Science
1988, Vol. 3, No. 4, 386–412

The Interface between Statistics and Philosophy of Science

I. J. Good

Abstract. Many points of contact between statistics and the philosophy of science are reviewed. Some of the topics mentioned are kinds of probability, weight of evidence (corroboration), tail probabilities (p-values), Bayes/non-Bayes compromises, "explicativity," induction and probabilistic causality.

Key words and phrases: Bayes factor, Bayes/non-Bayes compromise, explicativity, harmonic-mean rule of thumb, hierarchical Bayes, induction, penalized likelihood, probabilistic causality, standardized p-values, weight of evidence (corroboration).

INTRODUCTION

My topic is the interface between statistics and the philosophy of science; that is, the influence that each has had or might have on the other. Many people have contributed to this topic but I shall mainly review the writings of I. J. Good because I have read them all carefully. These influences are related to the semiquantitative ideas that emerge from an informal Bayesian approach, jestingly called Doogian. I don't want to repeat too much of what I have said in my books (Good, 1950, 1965, 1983f) and longer articles, but some overlap is necessary for the sake of intelligibility. I am less reticent about repeating what I have published in recent years in numerous short notes.

Among the topics that I shall touch upon are probability, surprise, rationality, corroboration or weight of evidence, explanation, induction, probabilistic causality and a Bayes/non-Bayes compromise.

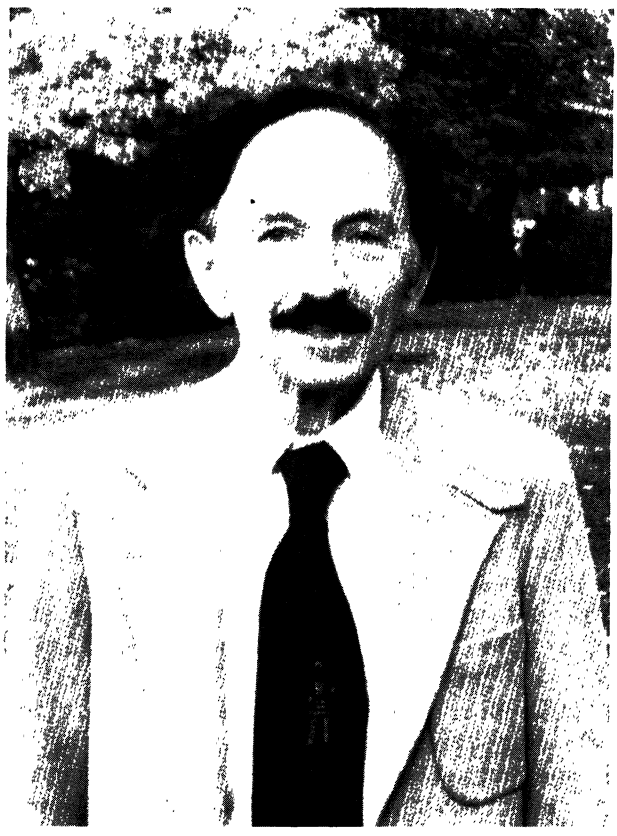
Philosophy goes beyond the dictionary in giving clearer meanings to the abstract words and phrases that have been found useful in ordinary language over the centuries. Sometimes more than one meaning is found and then the relations between them become of interest. The expressions "probability" and "weight of evidence" are two examples that have interested philosophers for two thousand years. "Probability" has several interpretations whereas "weight of evidence" has one that is clearly best in my opinion, once the interpretation of probability is chosen. Some philosophers are still using other definitions of weight of evidence without mentioning the best one.

My discussions belong to a field that can be called the mathematics of philosophy or probabilistic philosophy. The approach is often only semiquantitative because of the difficulty or impossibility of assigning precise numbers to the probabilities. Some people will argue that it is misleading to use precise-looking formulae for concepts that are not precise, but I think it is more leading than misleading because a formula encapsulates many words and provides a goal that one can strive toward by sharpening one's judgments. Also it is easier to make applications to statistics if one has a formula. A semiquantitative theory should be consistent with a good qualitative theory. For example, I think this applies basically to my theory of probabilistic causality (Good, 1961/62, 1985d, 1987a) in relation to the more qualitative theory of Suppes (1970). A reader who holds in mind the present paragraph will not be misled by the apparent precision of the formulae.

PROBABILITY

Poisson (1837, page 2) made a clear distinction between two kinds of probability which may be called epistemic and physical (see Good, 1986a, for further discussion). Epistemic probability can be either subjective (= personal) or logical and finer classifications have been given (Kemble, 1942; Good, 1959, 1966; Fine, 1973). Poisson assumes that the probability of an event is different for different people only because

I. J. Good is University Distinguished Professor of Statistics and Adjunct Professor of Philosophy, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. *This paper has been adapted from a talk presented by videotape at the Eighth International Congress of Logic, Methodology and Philosophy of Science in Moscow in August 1987, and which will appear in the proceedings of that conference.*



I. J. Good

they have different information. This seems to imply that $P(A \mid B)$ is the same for everybody so Poisson must have had credibility (= logical probability) in mind. Subjective probability was regarded as the most basic kind by Ramsey (1926/64) and de Finetti (1937/64) and in books by Good (1950) and Savage (1954). Early modern books on credibility were written by Keynes (1921), Jeffreys (1939) and Carnap (1950), although all three of these authors later became more sympathetic to the use of subjective probability than they were when they wrote those books.

I doubt whether credibility can ever be given a convincing precise numerical meaning unless the information has symmetry properties, or if the sample is very large, but I believe it is a useful fiction to assume that credibility has sharp values even when there is no sample, and I think it is mentally healthy for you to think of your subjective probabilities as estimates of credibilities. (The concept of useful fictions was developed by Jeremy Bentham in the early nineteenth century: see Ogden, 1959.) Physical probability too is a useful fiction even if the world is deterministic, just as pseudorandom numbers are regularly used by statisticians as if they were strictly random. (See both indexes of Good, 1983f, under "determinism.") de Finetti proved a theorem that can be interpreted as saying that a person who has sharp subjective probabilities that are consistent with the axioms behaves *as if* physical probabilities exist (although de Finetti believed they do not exist) and these physical probabilities have unique subjective probability distributions. The theorem can also be interpreted as saying that solipsism cannot be strictly disproved. (Solipsism is the theory that the only reality is one's own mind. There might be card-carrying solipsists but no sane sincere ones.) de Finetti did not express the theorem in either of these ways. It is an excellent example of a theorem at the interface between philosophy and statistics. For a simple exposition of de Finetti's theorem see Good (1965, pages 12–14, 22–23). For its relationship to the nondisprovability of solipsism see Good (1983f, pages 93 and 154), where further references are mentioned.

Keynes (1921) argued that credibilities should be regarded as interval valued, that is, partially ordered. Good (1950) adopted the same view for subjective probabilities, but sometimes it is a good enough approximation to think of a probability as having a sharp numerical value. The upper and lower subjective probabilities, which are the ends of the intervals, are also not strictly sharp, but again I believe it is often good enough to *assume* that they are sharp. I think the simplest satisfactory theory of partially-ordered subjective probability, or any other well-rounded scientific theory, is one based on axioms, rules of application and *suggestions*. I listed 27 suggestions in Good (1971, pages 124–127) and called them the Priggish Principles. But it would be too repetitive of what I've previously published to give the details here. See Good (1950, mostly written in 1947, 1962, 1982a, c, 1987c). In these works, I did not make quite explicit the fact that an exact additive measure cannot be ascribed to nonmeasurable sets. This was because I had in mind sets and propositions that correspond reasonably closely to definable events in the real world. No single nonmeasurable set is effectively constructible, but only by using the metaphysical axiom of choice.

In most circumstances your estimate of a probability should not be changed if you imagine that you are going to live forever (although the relevant utilities would change). Therefore it is legitimate to justify the axioms of subjective probability by reference to long-run frequencies. But there are more convincing justifications. The concept of the probability of a law of nature leads to difficulties for all approaches. In the frequency approach it requires us either to abandon the attempt or to imagine a large number of conceivable universes.

The use of partially-ordered probabilities can be regarded as a kind of "formalization of vagueness." It differs from the theory of fuzzy sets that deals with "degrees of belonging" to a set or, as it might be

expressed, with "degrees of meaning" (Good, 1950, page 1n). For example, it is more meaningful to say that a man has a beard if he resembles a religious leader, such as Christ, the Ayatollah, R. A. Fisher, Santa Claus or Karl Marx, than if his chin is merely fuzzy like one of the characters in Miami Vice. There might, however, be a correlation between the degree to which an entity x belongs to a specific set s and the probability that a person chosen at random would say that x belongs to s.

When all your prior probabilities are sharp you are a strict Bayesian, whereas, when all upper and lower point probabilities are 1 and 0, respectively, you are a strict non-Bayesian. Because I believe that subjective probabilities are only partially ordered I am forced into a Doogian intermediate position. I am forced to look for compromises between Bayesian and non-Bayesian methods and especially ways in which a somewhat Bayesian outlook can shed light on and improve so-called non-Bayesian methods. This point of view has been much developed in terms of mainstream statistics by Berger (1984).

I regard it as acceptable to use seemingly non-Bayesian methods *except when they are seen to contradict your own judgments of probabilities etc. in a given application*, the axioms of subjective probability being assumed. Whether you arrive at a contradiction will depend partly on how much thought you give to the matter. The type II principle of rationality recommends that you should allow for the cost of thinking and calculation when trying to apply the type I principle, namely the maximization of expected utility. Thinking will often cause you to change your mind; that is why dynamic probabilities are relevant: see, for example, Good (1977a).

There is a weak analogy between the concept of dynamic probability and the use made by Jeffrey (1965, pages 154 and 155) of two different notations for probabilities before and after some experience, but I am not convinced that two notations are required. The context of the problem is that of estimating a probability conditional on an uncertain event. My analysis of this problem is given in Good (1981c) where the connection with the logic of medical diagnosis is mentioned.

You can make probability judgments about the accuracy of your own judgments, and this leads to a hierarchical Bayesian approach in statistics, not necessarily restricted to only two levels. This approach is at least an aid to the judgment. It was exemplified by a so-called type II minimax procedure in a summer conference in Cambridge in 1951 (Good, 1952). Later it led to an adequate Bayesian significance test for multinomials (Good, 1965, 1967; Good and Crook, 1974; Leonard, 1977), and for contingency tables (Good, 1965, 1976a; Leonard, 1975; Crook and Good, 1980; Good and Crook, 1987). The basic idea is to use prior distributions that contain parameters known as hyperparameters, and these can be assigned hyperpriors. A history of much of the hierarchical Bayesian approach for categorical data, up to 1979, is given by Good (1980a). This work had nonhierarchical roots dating back to Bayes and Laplace, and, in this century to Johnson (1932) who anticipated some of the work of Carnap (1952). Johnson showed that, under certain assumptions, if one has a multinomial sample $(n_1, n_2, \ldots, n_t)$, then the physical probabilities of the t categories can be best estimated by adding a flattening constant k to all t observed frequencies. (His proof was incorrect for $t = 2$; see Good, 1965, page 26.) In Laplace's writings k was equal to 1. The flattening constant is a hyperparameter and I maintain that it should be assigned a hyperprior in order to improve the usefulness of the model.

The hierarchical Bayesian approach has also been applied to linear models especially by Lindley (1971) and Lindley and Smith (1972). A conference on hierarchical Bayesian statistics was held in 1986 in Bowling Green, Ohio, in which Smith was the principal speaker, and his lectures will be published.

My guess about the future of statistics is that it will be a compromise between hierarchical Bayesian methods and methods that seem superficially to be non-Bayesian.

INDUCTION

By scientific induction I mean changing the probability of hypotheses in the light of evidence or observations and thereby also changing the probabilities of future observations. The problem is partly solved by means of Bayes's theorem. Some people call the *formulation* of hypotheses "induction," but I prefer the obvious name *hypothesis formulation* for that activity. (Peirce called it "abduction.") Sometimes hypotheses can be formulated automatically by maximizing entropy: see Good (1963).

The estimation of physical probabilities of multinomial (or binomial) categories is of course a contribution to the problem of scientific induction. In particular the hierarchical Bayesian method was used explicitly for this purpose by Good (1983a,b). A qualitative consequence of the hierarchical approach, and of the calculations, was that "induction to the next trial" is much more reliable than "universal induction" or "induction to all future trials," and I think most people would agree with this conclusion without detailed analysis. I shall not give the details here because they are too mathematical and involve much

numerical computation. Instead I'd like to say something about another aspect of scientific induction.

The first quantitative contribution to scientific induction was Laplace's Law of Succession. For example, if you have seen n swans in England and they have all been white, and if you assume no other knowledge, then the odds are $n + 1$ to 1 that the next one chosen at random will be white according to Laplace's law, or $2n + 1$ to 1 on the basis of an "invariant" prior for the binomial parameter that was proposed by Jeffreys and independently by Perks (equivalent to using $k = \frac{1}{2}$ for binomials). If the conditions change in a substantial manner, for example, if the next observation is made in Australia, you cannot be so sure, and in fact there are *black* swans there. Similarly, after 24 successful launchings of the space shuttle Challenger, the odds that the next one would be successful would be 25 to 1 on or 49 to 1 on according to the two inductive procedures mentioned (but ∞ to 1 by maximum likelihood estimation). One engineer had estimated the chance of disaster as 1/35 for each of the previous launchings. But when the temperature on the next trial was 28° Fahrenheit whereas it had never been below 51° before, and when, for that reason, the engineers advised against launching (Hickey, 1987), the odds of a successful launching could hardly have been *rationally* estimated by those responsible as more than about 3 or 5 to 1 on, in my opinion, because the extra information seems to me to be worth a Bayes factor of about 10 against success. Of course this personal judgment might be described as "back-jobbing." I intend to compare it with judgments elicited from other people including "anti-Bayesians."

A special case of a hypothesis is that a specific word has a specific meaning or class of meanings, and this hypothesis is made more probable if you look the word up in a dictionary and also observe how the word is used. This applies to every word in the language including "induction" itself, so if some one tells me he doesn't believe at all in probabilistic induction, for all I know he is asserting that the moon is made of gorgonzola or that pigs eat purple people (Black, 1967; Good, 1981b). It is like a nondreaming solipsist trying to convince other people he is right. Popper and Miller (1983) produced a new argument against probabilistic induction, and because I was sure their conclusion was wrong, if I understood what they meant by probabilistic induction, I seized on what I thought was the weakest link of their argument in my first response. But it turned out that that was not a necessary link in their argument, so I'm forced to the opinion that their argument was a kind of ingenious sleight of hand or a *non sequitor*. (See Good, 1985a, for more details.) Their argument was attacked by Redhead (1985) and then Redhead was seemingly refuted by Gillies (1986). I reinstated Redhead's argument, at least temporarily, by replacing his implicit definition of support or weight of evidence by the best definition (Good, 1987b). (But, at the time of writing, Popper and Miller (1987) have had the last word.) I shall discuss weight of evidence next.

WEIGHT OF EVIDENCE

The earliest use of the expression "weight of evidence" quoted by the *Oxford English Dictionary* (1971) is a remark made by T. H. Huxley in 1878. In that same year Peirce (1878) published a formal definition, so either the expression was already in common use or perhaps it is simply self-explanatory in a qualitative sense. Clearly the concept was familiar to the ancient Greeks because Themis, the goddess of justice, is said to be represented as holding a pair of scales in which she weighs opposing arguments. I shall now outline an argument that leads to a unique explicatum. I have discussed the topic of weight of evidence in over forty publications beginning with Good (1950) and there are surveys (Good, 1985c, 1988c). The desideratum-explicatum argument outlined in the present text was given in Good (1968b) and much more lucidly in Good (1984f).

Let H denote a hypothesis, such as that an accused person is guilty, and let E denote some evidence, such as that presented by a specific witness. We ask how should we define $W(H: E \mid G)$, the weight of evidence in favor of H *provided by* E when background knowledge G is regarded as given or previously taken into account. It is natural to assume that the new evidence converts the prior probability into its posterior probability, that is, that $P(H \mid E\&G)$ is a mathematical function of $P(H \mid G)$ and of the weight of evidence. Moreover, $W(H: E \mid G)$ should depend only on (i) the probability of E given that the accused is guilty, and (ii) the probability of E given that he is innocent, that is on $P(E \mid H\&G)$ and $P(E \mid \bar{H}\&G)$ where the bar denotes negation. These desiderata lead to the conclusion that $W(H: E \mid G)$ must be a monotonic function of the *Bayes factor* $P(E \mid H\&G)/P(E \mid \bar{H}\&G)$ and we may as well take the logarithm of the Bayes factor as our explicatum because this leads to desirable additive properties of the kind assumed by the goddess Themis. In fact,

$$(1) \quad W[H: (E\&F)] = W(H: E) + W(H: F \mid E).$$

I have taken G for granted to simplify the appearance of the formula. When E and F are independent given H and also given $\bar{H}$, this formula reduces to

$$W(H: E\&F) = W(H: E) + W(H: F).$$

It was pointed out by Wrinch and Jeffreys (1921), in a slightly different notation, that

$$\frac{P(E \mid H\&G)}{P(E \mid \bar{H}\&G)} = \frac{O(H \mid E\&G)}{O(H \mid G)}, \tag{2}$$

the ratio of the final (posterior) to the initial (prior) odds. Thus, W is the additive change in the log odds of H by virtue of E. It would be a misuse of terminology, and also historically misleading, to call the left side of (2) a likelihood ratio although it *is* a likelihood ratio when H and $\bar{H}$ are what are called simple statistical hypotheses. It would be acceptable to call it a Bayesian likelihood ratio, but there is then a risk that the qualification "Bayesian" would ultimately be dropped, thus leading back to the misleading terminology. Furthermore, "likelihood ratio" is often used to mean the ratio of *maximum* likelihoods. Equation (2) has been mentioned several times in the literature without citing Wrinch and Jeffreys. Because it is so important I think proper credit should be given although it is an easy deduction from Bayes's theorem. Bayes's theorem is an easy deduction from the axioms of epistemic probability, but authors don't write it down as if they had just discovered it for the first time.

It is best to think of the Bayes factor as defined by the right side of equation (2), that is, as the factor by which the initial odds of H are multiplied to obtain the final odds. It is convenient that this factor is equal to the left side because this can be evaluated independently of the initial probability of H which can be especially difficult to judge. I conjecture that most juries are able to judge *final* probabilities of guilt better than *initial* probabilities, because in ordinary affairs final probabilities are more important than initial ones so we think about them more.

Because the left side of (2) sometimes reduces to a simple likelihood ratio we can regard a Bayes factor as part of the interface between Bayesian and less philosophical non-Bayesian statistics. Ordinary (non-Bayesian) likelihood is also part of this interface.

The technical concept of weight of evidence, because it captures the intuitive concept so well, should be of interest in legal matters (Good, 1986c), and is already of interest for medical diagnosis, especially differential diagnosis (between two diseases): see, for example, Good and Card (1971), Card and Good (1974) and Spiegelhalter and Knill-Jones (1984).

The concept of a *unit* of weight of evidence is due to Turing (1941). He talked of bans, decibans and natural bans, the latter when natural logarithms are used. The deciban resembles the decibel in acoustics, being about the smallest weight of evidence perceptible to the human mind. Turing's name for a weight of evidence was "score" or "decibannage."

Peirce (1878) (long before Fisher introduced the technical meaning of "likelihood," indeed twelve years before Fisher was born) almost anticipated the best formal concept of weight of evidence but his definition applies only if the initial odds of H are 1 or "evens," that is, if $P(H \mid G) = \frac{1}{2}$. In this special case, the weight of evidence is equal to the posterior log odds. Jeffreys (1939) also nearly always assumes that $O(H) = 1$ in spite of his earlier work. This was because in his book he was trying to be a credibilist, especially in the 1939 edition. Poisson (1837, Chapter V) also came close to the formal concept: see Good (1986a, page 167).

Weight of evidence can be regarded as a quasi-utility or epistemic utility, that is, as a substitute for utility when the actual utilities are difficult to estimate. (A quasi-utility can be defined as an additive epistemic utility.) Just as for money, diminishing returns eventually set in; for example, in a court of law, if the weight of evidence in favor of guilt or innocence becomes overwhelming there is little point in seeking further evidence, especially if it is expensive. The same principle applies in scientific or medical research or even in a game of chess (where evolving or dynamic probabilities are relevant: see Good, 1968b, and especially 1977a). But the effect of diminishing returns can often be ignored. When this is done we naturally bring in the concept of *expected* weight of evidence, which, in discriminating between two multinomials, leads to an expression of the form

$$\sum_i p_i \log(p_i/q_i). \tag{3}$$

This, or its general form (continuous or mixed), is often called cross-entropy, or relative (neg)entropy. Such expressions were used by Gibbs (1875/1906/1961, page 163), somewhat implicitly, in statistical mechanics and in statistics by a number of authors. For many references see Good (1985c), Christensen (1983, Chapter 1) and the indexes of Good (1983f) under "weight of evidence." Ordinary entropy is effectively minus a special case of cross-entropy, namely when q_i has the same value for all i. In the design of an experiment for estimating a parameter it might be reasonable to maximize the expected cross-entropy; but to minimize the cross-entropy when doing the estimation after the experiment is done. (Compare Good, 1968a.) This is because, according to a theorem due to Wald (1950, page 18), a minimax solution is a Bayes solution that uses the *least* favorable prior. Minimax solutions are not optimal but they have the merit of invariance under changes of variables. (See also Good, 1955/56, 1969; Lindley, 1956.) It seems that whole areas of statistics can be regarded as having their logical roots in the concept of weight of evidence and its mathematical expectation. This is

not surprising because weight of evidence is a concept almost as fundamental as probability itself. I believe that even Shannon's coding theorems in communication theory are understood better in terms of expected weight of evidence rather than entropy (Good and Toulmin, 1968).

TAIL PROBABILITIES OR *P*-VALUES

In statistical practice a small p-value such as $1/50$ is usually regarded as evidence against the "null hypothesis" H, and there is a temptation to think that any fixed value, say $p = 0.031$ (which is not at all the same assertion as that $p < 0.05$; see Good (1950), page 94n) conveys the same amount of evidence against H on all occasions, at any rate if we are careful to use either single tails or double tails, depending on circumstances. (Perhaps Fisher (1956), page 100, implies this view, for example, when he refers to "the weight of evidence.") This temptation must be resisted for several different totally convincing reasons. Some of these reasons are mentioned in my paper on hypothesis testing (Good, 1981a), and I shall not repeat them here. Here I'd like to mention that a very simple argument can be given, without mentioning Bayes or Neyman and Pearson, to prove conclusively the diminishing significance of a fixed p-value when a sample size is increased (Good, 1983c). Indeed, *given a fixed statistical model*, a fixed p-value, however small, can *support* the null hypothesis if the sample size is large enough and if the mathematical model is *sufficiently reliable*. Many statisticians are still surprised to hear that this conclusion is true even if the null hypothesis is absolutely sharp. But in practice a small "neighborhood" should usually in principle be included around the null hypothesis (as pointed out by Laplace in 1774 according to Stigler (1986, page 135n) and somewhat later independently by Good (1950, pages 90–93). When this is done it is obvious that a small p-value with respect to the *sharp* null hypothesis might support the "enlarged" null hypothesis. (For a brief discussion of what it means to say that a theory is *true* see Good, 1986f.)

In several situations the Bayes factor against a sharp null hypothesis is roughly proportional to $1/(P\sqrt{N})$: see Jeffreys (1939, Appendix 1), Good (1983f, page 143). One way to understand this is that the prior measures of reasonable sets of non-null hypotheses, such as 97½% confidence intervals, shrink roughly proportionally to $1/\sqrt{N}$. I have accordingly suggested (Good, 1982b, 1984a, g, h) that p-values, if you *must* use them, should be standardized to a fixed sample size, say $N = 100$, by replacing P by

$$\min(\tfrac{1}{2},\ P\sqrt{N/100}) \tag{4}$$

(when $N > 10$) and calling it a p-value *standardized to sample size* 100. The reason for the ½ is given in the cited reference. Even a fixed standardized p-value does not correspond to the same weight of evidence for all occasions, but it is better in this respect than an ordinary p-value. I guess that standardized p-values will not become standard before the year 2000. Of course if you are *sure* that there are only two simple statistical hypotheses, then there is little point in using p-values instead of Bayes factors.

Standardized p-values exemplify the concept of a Bayes/non-Bayes compromise. Several other examples, and historical comments, can be found, for example, in a recent encyclopedia article on scientific method and statistics (Good, 1988a). One example is the concept of the *strength* of a test (Crook and Good, 1982).

For some recent discussion of p-values see Berger and Sellke (1987), Casella and Berger (1987) and Good (1986e).

THE COMBINATION OF *P*-VALUES IN PARALLEL

Let $P_1, P_2, \cdots$ be some p-values obtained by distinct tests, but *based on the same data*. I call these "tests in parallel." A dishonest experimenter might choose the smallest or largest of these depending on whether he is bribed or intimidated to disprove or to support the null hypothesis. A rule of thumb that seems to appeal even to non-Bayesians is to replace these p-values by their harmonic mean or perhaps by a weighted harmonic mean. This proposal has an informal Bayesian justification, and is a nice example of a Bayes/non-Bayes compromise (Good, 1958). The argument was based on the fact that a Bayes factor against a null hypothesis is often, in any given experiment of fixed sample size, very roughly proportional to $1/p$, at least when p is less than say ¼. There have been recent elaborations and applications of this harmonic mean rule of thumb; Good (1984b,c,d,e). For example, you might not know whether a comparative experiment should be regarded as paired or unpaired and the two possibilities would give two different p-values that could be combined by the rule of thumb. The rule of thumb was published 29 years ago in a well-known periodical, but is still being ignored because most statisticians like to pretend their methods are precise. I think it is better to be approximately correct than precise and wrong. Perhaps this will be the usual opinion by the year 2001.

THE CHOICE OF A CRITERION FOR A SIGNIFICANCE TEST

An early example of a Bayes/non-Bayes compromise, understood explicitly as such, was related to the

choice of a criterion for a significance test (Good, 1957, page 863). The proposal was to compute a Bayes factor (or equivalently a weight of evidence), based on a Bayesian model in which you do not necessarily have much confidence, and then to treat this Bayes factor merely as a criterion in the "Fisherian" manner so to speak by obtaining its distribution given the null hypothesis. (I sometimes refer to the use of p-values as Fisherian to distinguish this usage from the acceptance-rejection procedure of Neyman and Pearson and from strict Bayesian methods. But p-values have a history dating back for two centuries. The idea of finding the distribution of a Bayes factor, given the null hypothesis, is now practicable up to a point by simulation (Good, 1986d).) Fisher used to select criteria for significance tests without any explicit formal principle, but based on "common sense," although he had explicit principles for estimation problems. His common sense was undoubtedly based on some vague non-null composite hypotheses, in fact he said (Fisher, 1955, page 73), in relation to p-values, that "The deviation [might be] in the direction expected for certain influences *which seemed to me* not improbable ..." (my italics). Note the personal Bayesian tone here and on page 74 he refers to "the tester's state of mind." I wonder where this explicit subjectivism first occurred in Fisher's writings. Did it occur before the revival of the modern subjectivistic movement? (Soon after Good (1950) was published Fisher told a common colleague that he had found it interesting. Giants can be influenced by the dwarfs that stand on their shoulders and whisper in their ears.) Strict Bayesians and Neyman-Pearsonians have to select precise non-null hypotheses, although in reality there is nearly always some vagueness in the real world. How much should be formalized and how much should be left vague depends partly on personal judgment.

Note that the Neyman-Pearson-Wilks "likelihood ratio," a ratio of maximum likelihoods, can be regarded as a crude approximation to a Bayes factor for a *very bad* Bayesian model, yet it works well as a significance criterion. The basic idea is that, if you cannot evaluate an integral, work instead with the maximum of the integrand without even allowing for the curvature of the integrand at its maximum! This crude idea also leads to the use of maximum likelihood as another example of a Bayes/non-Bayes compromise.

When the number of parameters is large this informal Bayesian justification of maximum likelihood estimation is liable to break down, and then, I believe, the method of maximum likelihood becomes unacceptable. A very good example is the estimation of a probability density function f given a finite sample of observations $x_1, x_2, \cdots, x_N$. In this case the number of *parameters* is infinite and the maximum likelihood estimate consists merely of one Nth of a Dirac function at each observation. This disaster can be avoided by using the method of maximum penalized likelihood in which the log likelihood $\sum_i \log f(x_i)$ is penalized by subtracting from it a *roughness penalty* $\phi(f)$ such as $\beta \int [(\sqrt{f})'']^2\, dx$ where β is called a hyperparameter or smoothing parameter (Good and Gaskins, 1971, 1972, 1980; Good and Deaton, 1981; Leonard, 1978). The method of maximum penalized likelihood was described by Good and Gaskins (1972) as a wedding between Bayesian and non-Bayesian methods because one can either regard $\exp(-\phi)$ as proportional to a prior density (possibly improper) in function space or else the whole procedure can be regarded as a common sense *ad hoc* non-Bayesian adjustment of maximum likelihood estimation to save it from disaster. (A special feature of the Bayesian interpretation is that it leads to a way of evaluating bumps.) A similar penalizing of a log likelihood was also suggested by Good (1963, page 931). The idea proposed there, but not developed, was to maximize a linear combination of log likelihood and entropy. When there is no sample this suggestion reduces to the method of maximum entropy, so the proposal was a generalization of that method.

For all these procedures it is necessary to choose the hyperparameter, procedural parameter or smoothing parameter. Methods are given by Good and Gaskins but their reliability needs to be investigated by further simulation methods. One could also assume a hyperprior for the hyperparameter. If a sample is large then the smoothing parameter can be reliably estimated by using the old-fashioned split sample method or by means of the modern modifications called cross-validation or predictive sample reuse, although these methods can be expensive. For further discussion of the cross-validation method for density estimation techniques see Wahba (1977).

The theory of significance tests, based on p-values, cannot be entirely separated from the theory of estimation of parameters. Thus, Fisher (1955) said "... in the theory of estimation we consider a continuum of hypotheses each eligible as a null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turn as true—including frequencies of error [p-values], therefore only of the 'first kind,' without any assumptions of knowledge *a priori*—which supply the likelihood function, fiducial limits, and other indications of the amount of information available." In this way Fisher was able to subsume the concept of errors of the second kind under those of the first kind. This p-value function is a continuous form of all possible confidence intervals, although Fisher might have deliberately avoided this mode of

expression! It is not surprising that Pearson (1955) said, in response, that "· · · I do not think that our position in some respects was or is so very different from that which Professor Fisher himself has *now* reached" (my italics). Another implication of Fisher's remark is in suggesting the notion of a continuum of hypotheses possibly forming an "onion," or part of an onion, surrounding the null hypothesis, these hypotheses being "more non-null" when further out, and the inner core being virtually the null hypothesis.

Fisher mentioned fiducial limits in the quoted passage, so I'll remind you *en passant* that, in my opinion, Fisher's fiducial argument is fallacious and the reason he made a mistake was simply because he did not use a notation for conditional probability. In Fisher (1955, page 24) he says "He [Neyman] seems to claim that the statement (a) 'θ has a probability of 5 per cent of exceeding T' is a different statement from (b) 'T has a probability of 5 per cent of falling short of θ'." In my opinion the error was Fisher's because only one of the statements should be made conditional on T. Bad notations and terminology tempt people into making substantial errors. An example was Carnap's use of "confirmation" for logical probability, a usage that still causes confusion among philosophers of science. The ordinary English meaning of confirmation is much closer to weight of evidence than to probability. I predict that the misuse of the term "confirmation" will continue until the year 2002.

Jeffreys (1939) showed that in some circumstances the use of the fiducial argument was equivalent to assuming a specific Bayesian prior, usually "improper," that is, integrating to infinity instead of 1. According to Stigler (1986, pages 91 and 102–104) the fiducial argument was foreshadowed by Thomas Simpson in 1755 and its relation to inverse probability was recognized, but only implicitly, by Laplace. The error in the exposition of the fiducial argument by Fisher (1956), *together with the psychological reason for the error*, has been precisely pinpointed (Good, 1971, page 139).

SURPRISE INDEXES

A kind of alternative to the use of p-values are surprise indexes. To save space I refer you to a review of this topic in Good (1988b).

PROBABILISTIC CAUSALITY

Sometimes "causality" is taken to mean "determinism" as when people say that quantum mechanics sounded the death knell for causality. In the present context it is convenient to refer to determinism as *strict causality* and to refer to something less strict as *probabilistic causality*. Work on quantum mechanics during the last decade seems to refute determinism but it also seems to refute objective reality (d'Espagnat, 1979) so, following Einstein, me-thinks the theory refutes too much.

If the world is deterministic then probabilistic causality does not exist, but we'll never know with certainty whether determinism or indeterminism is true. So it is legitimate to assume indeterminism even if it is only a convenient fiction, somewhat like using the axiom of choice in a mathematical proof. There would be no criminal law if, believing in determinism, we always said "Tout comprendre c'est tout pardonner." Anyway *nous ne tout comprendon jamais*, we never understand everything.

It is essential to make a distinction between the tendency of one event F to cause a later one E, denoted by $Q(E:F)$, and the extent to which F actually caused E, denoted by $\chi(E:F)$. I gave a convincing example of this in Good (1961/62) based on a dramatic incident involving Sherlock Holmes, Watson and Moriarty. In the law, a simple example is the distinction between murder and attempted murder. The distinction is important because the law rewards inefficiency in this case, at least in many countries.

The notations $Q(E:F)$ and $\chi(E:F)$ are both only abbreviations because one must allow also for the state U of the universe just before F occurred and also for all true laws of nature. It is also necessary to allow for the negations of E and of F but when you put all these aspects into the notation in a lecture some people walk out because they think you are doing mathematics.

It is extremely difficult to find a fully satisfactory explicatum for χ, although I think I have made some contribution toward it. (For my latest effort see my reply to a valid criticism by Salmon in Good, 1987a.) Here I shall discuss only Q which seems to me to be of much greater importance in statistics, although in legal matters χ is at least as important. It will be Q that counts when you reach "dem pearly gates." I won't say much even about Q because I have recently given two lectures on the topic (Good, 1985d, 1987a).

The old-fashioned name "the probability of causes" referred to the application of Bayes's theorem, where the "hypotheses" are regarded as mutually exclusive possible "causes" of some event or events. For example, the "event" might be a set of medical indicants and the possible "causes" might be various disease states. The topic I'm discussing now is different: it refers to the tendency of some event F to cause another one E, not the probability that F was *the* cause of E.

Let us assume that $Q(E:F)$ is some function of all probabilities of the form $P(A \mid B)$ where A and B are logical combinations of E and F. This comes to the same thing as assuming that Q depends only on $P(E \mid F)$, $P(E \mid \bar{F})$ and $P(F)$. The probabilities are

here best regarded as physical (propensities) because I am thinking of probabilistic causality as something that exists even if no conscious being is around. By assuming several desiderata related to the "causal strengths" and "causal resistances" of causal networks, we can arrive at the explicatum that Q is equal to the weight of evidence against F provided by the nonoccurrence of E; that is,

$$(5) \quad Q(E:F) = W(\bar{F}:\bar{E}) = \log\left[\frac{1 - P(E \mid \bar{F})}{1 - P(E \mid F)}\right]$$

(where of course U is taken for granted throughout). This expression is mathematically independent of $P(F)$ and this could have been taken as a desideratum although it was not explicitly used to obtain the explicatum. That $Q(E:F)$ is mathematically independent of $P(F)$, the initial probability of F, is desirable for the following reason.

In a scientific experiment we might decide whether to apply a treatment F by using a randomizing device that would determine $P(F)$. The purpose of the experiment might be to find out to what extent F causes E by repeating the experiment many times. It would be contrary to the spirit of scientific experimentation if the conclusion were to depend on our arbitrary choice of $P(F)$. Some people would go further and would say that no reliable conclusions are possible unless the experimenter uses a randomizing device to control whether F occurs. In this way we can be convinced that E and F did not have a common cause unless we believe in some possibly magical or paranormal effect that relates the randomizing device to the effectiveness of the treatment. This is why it is reassuring to discover that the proposed explicatum for Q does not depend on $P(F)$, although this property was not used in the original derivation of the explicatum.

It seems intuitively right that Q should have something to do with weight of evidence. So what happens if we define $Q(E:F)$ by some other weight of evidence, the possibilities being (i) $W(F:E)$, (ii) $W(E:F)$ and (iii) $W(\bar{E}:\bar{F})$? The second and third possibilities can be excluded because they depend on the initial probability of F, $P(F)$. So the only rival to $W(\bar{F}:\bar{E})$ is $W(F:E)$, still conditional on U of course. This rival will now be ruled out. Consider the "game" of Russian roulette (possibly called American roulette in Moscow). In a self-explanatory notation, and for an obvious slightly oversimplified model, we have $P(E \mid F) = 1/6$, $P(E \mid \bar{F}) = 0$, if the "game" is played with a six shooter that contains just one bullet. Hence, $W(\bar{F}:\bar{E}) = \log(6/5) = 78$ centibans (or "centicausits"), whereas $W(F:E) = \log[5/6/0] = \infty$. It makes sense that a necessary cause of E should have only a finite tendency to cause E, whereas a sufficient cause should have an infinite tendency if E was not already inevitable. Playing Russian roulette is a necessary but not a sufficient cause for disaster in the assumed model. Similarly, trying to cross the road is usually a necessary cause for getting run over, but fortunately it is not sufficient. Thus, $W(F:E)$ is shot down. We see then that if $Q(E:F)$ is to be expressed in terms of weight of evidence there is really only one serious candidate, namely $W(\bar{F}:\bar{E})$. Moreover this has desirable additive properties (for example, Good, 1983f, page 209) that would not be shared by any function of it other than a mere multiple.

It turns out that $Q(E:F)$, as thus explicated, is identical with one of the measures of association used for 2 by 2 contingency tables. Also there is a relationship to the theory of linear regression, but I'll just give you citations for this (Good, 1980b, 1985d, 1987a).

EXPLICATIVITY

Popper (1959) suggested that a measure or index of explanatory power should be developed and this was a main theme of, for example, Good (1968b). I there introduced the concepts of weak and strong explanatory power and gave one statistical example. The qualification "strong" means that a penalty is paid for cluttering a hypothesis with irrelevances. I returned to the topic in Good (1977b) where the name strong explanatory power was changed to explicativity.

By explicativity η is meant the extent to which one proposition or event explains why another should be believed, to express the matter a little too briefly. The concept is not intended to capture all the senses of "explanation." A desideratum-explicatum approach was used leading quickly to the explicatum

$$(6) \quad \begin{aligned} &\eta(E:H) \\ &\quad = \log P(E \mid H) - \log P(E) + \gamma \log P(H) \end{aligned}$$

where $0 < \gamma < 1$ and where $\gamma = 1/2$ might be adequate. We can think of γ as a clutter constant because the more we object to cluttering H with irrelevancies, the larger we would make γ.

The amount by which the explicativity of H exceeds that of H' is

$$(7) \quad \begin{aligned} &\eta(E:H/H') \\ &\quad = (1 - \gamma)W(H/H':E) + \gamma \log O(H/H' \mid E), \end{aligned}$$

a compromise between the weight of evidence provided by E on the one hand, and the posterior log odds on the other hand. If we take $\gamma = 1$ there is no better hypothesis than a tautology such as $1 = 1$. If we take $\gamma = 0$ we ignore the prior probabilities. We must therefore compromise.

If explicativity is regarded as a kind of quasi-utility, its maximization leads to a method for choosing

among hypotheses, and this principle can be used in statistical problems of both estimation and significance testing. The results of applying this method make intuitive sense in several examples. For example, the method leads to interval estimation of parameters in these examples without assuming in advance that interval estimates should be used. The method is very general and could be used, for example, for the selection of regressor variables. The result would resemble methods proposed by Akaike (1974) and by Schwarz (1978). The maximization of *expected* explicativity is a reasonable recipe for experimental design, and it can be seen that γ then becomes irrelevant and the method reduces to that cited soon after equation (3).

The notion of explicativity seems appropriate for a semiquantitative discussion of how good natural selection is as an explanatory theory as compared with other theories of evolution (Good, 1986b).

For more on explicativity see Good and McMichael (1984) and Good (1985b).

ADHOCKERY

When a hypothesis or theory H appears to be undermined by the total relevant evidence E a defender of H might patch it up by changing it to a more elaborate hypothesis H'. Then has H been improved or is the change merely *ad hoc*? The concept of explicativity provides at least a formal solution to this problem: the change is *ad hoc* if $\eta(E:H/H')$ is positive, and $\eta(E:H/H')$ is a measure of the adhockery. If it is negative then the change is justified (compare Good, 1983d).

"SCIENTIFIC METHOD"

Somewhat supplementary to what I have said in this lecture is an encyclopedia article entitled "Scientific method and statistics" (Good, 1988a). In that article I tried to define scientific method in terms of fourteen facets and to argue that statistics makes use of all of these facets. This does not show that statistics is identical with the scientific method but only that statistics is one example of the method. For each way of assigning weights to the facets one gets a different interpretation of "scientific method."

EXPLORATORY DATA ANALYSIS

At first it might seem that exploratory data analysis is nonphilosophical but I believe it has implicit Bayesian aspects. This is argued in Good (1983e).

TECHNIQUE VERSUS PHILOSOPHY

Because I have been emphasizing the interface between philosophy and statistics, I might have given the impression that statistics is nothing but philosophy. That has not been my intention. Much of statistics consists of techniques for condensing data sets into simplified numerical and graphical forms that can be more readily apprehended by the eye-brain system, a system that has evolved at a cost of some 10^{18} organism-hours. Philosophers recognize the importance of techniques and technicians should reciprocate.

ACKNOWLEDGMENT

This work was supported in part by a grant from the National Institutes of Health.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

BERGER, J. O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analyses* (J. B. Kadane, ed.) 63–124. North-Holland, Amsterdam.

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilibility of P values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–133, 135–139.

BLACK, M. (1967). Induction. In *The Encyclopedia of Philosophy* **4** 169–181. Macmillan and The Free Press, New York.

CARD, W. I. and GOOD, I. J. (1974). A logical analysis of medicine. In *A Companion to Medical Studies* (R. Passmore and J. S. Robson, eds.) **3** Chapter 60. Blackwell, Oxford.

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

CARNAP, R. (1952). *The Continuum of Inductive Methods*. Univ. Chicago Press, Chicago.

CASELLA, G. and BERGER, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82** 106–111, 123–135.

CHRISTENSEN, R. (1983). *Multivariate Statistical Modelling*. Entropy Limited, Lincoln, Mass.

CROOK, J. F. and GOOD, I. J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. *Ann. Statist.* **8** 1198–1218.

CROOK, J. F. and GOOD, I. J. (1982). The powers and "strengths" of tests for multinomials and contingency tables. *J. Amer. Statist. Assoc.* **77** 793–802.

DE FINETTI, B. (1937/64). Foresight: its logical laws, its subjective sources. Translated from the French of 1937 by H. Kyburg, in *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds.) 95–158. Wiley, New York. Corrected in 2nd ed., 1980, 55–118. Krieger, Huntington, N. Y.

D'ESPAGNAT, B. (1979). The quantum theory and reality. *Sci. American* **241** 158–181.

FINE, T. L. (1973). *Theories of Probability*. Academic, New York.

FISHER, R. A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc. Ser. B* **17** 69–78.

FISHER, R. A. (1956). *Statistical Methods and Scientific Evidence*. Oliver and Boyd, Edinburgh.

GIBBS, J. W. (1875/1906/81). On the equilibrium of heterogeneous substances. In *The Scientific Papers of J. Willard Gibbs* **1**. Longmans, Green, London. Reprinted by Dover, New York, 1961, 55–349.

GILLIES, D. (1986). In defense of the Popper-Miller argument. *Philos. Sci.* **53** 110–113.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin, London.

GOOD, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14** 107–114. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1955/56). Some terminology and notation in information theory. *Proc. IEE C* **103** 200–204. Also Monograph 155R (1955).

GOOD, I. J. (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28** 861–881.

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Assoc.* **53** 799–813.

GOOD, I. J. (1959). Kinds of probability. *Science* **129** 443–447. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1961/62). A causal calculus. *British J. Philos. Sci.* **11** 305–318; **12** 43–51; **13** 88. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1962). Subjective probability as the measure of a non-measurable set. In *Logic, Methodology, and Philosophy of Science* (E. Nagel, P. Suppes and A. Tarski, eds.) 319–329. Stanford Univ. Press, Stanford, Calif. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, Mass.

GOOD, I. J. (1966). How to estimate probabilities. *J. Inst. Math. Applications* **2** 364–383.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. Ser. B* **29** 399–431.

GOOD, I. J. (1968a). Some statistical methods in machine-intelligence research. *Virginia J. Sci.* **19** 101–110.

GOOD, I. J. (1968b). Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. *British J. Philos. Sci.* **19** 123–143.

GOOD, I. J. (1969). What is the use of a distribution? In *Multivariate Analysis II* (P. R. Krishnaiah, ed.) 183–203. Academic, New York.

GOOD, I. J. (1971). The probabilistic explication of information, evidence, surprise, causality, explanation, and utility (with discussion). In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 108–141. Holt, Rinehart and Winston, Toronto.

GOOD, I. J. (1976a). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4** 1159–1189.

GOOD, I. J. (1977a). Dynamic probability, computer chess, and the measurement of knowledge. In *Machine Intelligence* (E. W. Elcock and D. Michie, eds.) **8** 139–150. Ellis Horwood, Chichester. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1977b). Explicativity: a mathematical theory of explanation with statistical applications. *Proc. Roy. Soc. London Ser. A* **354** 303–330. (Erratum (1981) **377** 504.) Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1980a). Some history of the hierarchical Bayesian methodology (with discussion). In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 489–510, 512–519. Univ. Press, Valencia, Spain. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1980b). Degrees of causation in regression analysis. *J. Statist. Comput. Simulation* **11** 153–155.

GOOD, I. J. (1981a). Some logic and history of hypothesis testing. In *Philosophy in Economics: University of Western Ontario Series on the Philosophy of Science* (Joseph C. Pitt, ed.) 149–174. Reidel, Dordrecht. Reprinted in *Good Thinking* (1983).

GOOD, I. J. (1981b). Can scientific induction be meaningfully questioned? *J. Statist. Comput. Simulation* **13** 154.

GOOD, I. J. (1981c). The weight of evidence provided by uncertain testimony or from an uncertain event. *J. Statist. Comput. Simulation* **13** 56–60.

GOOD, I. J. (1982a). The axioms of probability. In *Encyclopedia of Statistical Sciences* **1** 169–176. Wiley, New York.

GOOD, I. J. (1982b). Standardized tail-area probabilities. *J. Statist. Comput. Simulation* **16** 65–66.

GOOD, I. J. (1982c). Degrees of belief. In *Encyclopedia of Statistical Sciences* **2** 287–293. Wiley, New York.

GOOD, I. J. (1983a). The robustness of a hierarchical model for multinomials and contingency tables. In *Scientific Inference, Data Analysis, and Robustness* (G. E. P. Box, Tom Leonard and Chien-Fu Wu, eds.) 191–211. Academic, New York.

GOOD, I. J. (1983b). Scientific induction: universal and predictive. *J. Statist. Comput. Simulation* **16** 311–312.

GOOD, I. J. (1983c). The diminishing significance of a fixed P-value as the sample size increases: a discrete model. *J. Statist. Comput. Simulation* **16** 312–314.

GOOD, I. J. (1983d). A measure of adhockery. *J. Statist. Comput. Simulation* **16** 314.

GOOD, I. J. (1983e). The philosophy of exploratory data analysis. *Philos. Sci.* **50** 283–295.

GOOD, I. J. (1983f). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press, Minneapolis, Minn.

GOOD, I. J. (1984a). How should tail-area probabilities be standardized for sample size in unpaired comparisons? *J. Statist. Comput. Simulation* **19** 174.

GOOD, I. J. (1984b). One tail versus two tails, and the harmonic-mean rule of thumb. *J. Statist. Comput. Simulation* **19** 174–176.

GOOD, I. J. (1984c). Paired versus unpaired comparisons and the harmonic-mean rule of thumb. *J. Statist. Comput. Simulation* **19** 176–177.

GOOD, I. J. (1984d). The harmonic-mean rule of thumb; some classes of applications. *J. Statist. Comput. Simulation* **20** 176–179.

GOOD, I. J. (1984e). A sharpening of the harmonic-mean rule of thumb for combining tests "in parallel." *J. Statist. Comput. Simulation* **20** 173–176.

GOOD, I. J. (1984f). The best explicatum for weight of evidence. *J. Statist. Comput. Simulation* **19** 294–299; **20** 89.

GOOD, I. J. (1984g). Standardized tail-area probabilities: a possible misinterpretation. *J. Statist. Comput. Simulation* **19** 300.

GOOD, I. J. (1984h). The tolerant Bayesian's interpretation of a tail-area probability. *J. Statist. Comput. Simulation* **19** 300–302.

GOOD, I. J. (1985a). Probabilistic induction is inevitable. *J. Statist. Comput. Simulation* **20** 323–324.

GOOD, I. J. (1985b). Explanatory power depends on more than probabilities. *J. Statist. Comput. Simulation* **22** 184–186.

GOOD, I. J. (1985c). Weight of evidence: a brief survey (with discussion). In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 249–269. North-Holland, Amsterdam.

GOOD, I. J. (1985d). Causal propensity: a review. *PSA 1984* (P. D. Asquith and P. Kitcher, eds.) **2** 829–850. Philosophy of Science Association, East Lansing, Mich.

GOOD, I. J. (1986a). Some statistical applications of Poisson's work (with discussion). *Statist. Sci.* **1** 157–180.

GOOD, I. J. (1986b). Neo"-Darwinism. *Physica* **22D** 13–30. Also in *Evolution, Games, and Learning* (D. Farmer, A. Lapedes, N. Packard and B. Wendroff, eds.) 13–30. North-Holland, Amsterdam.

GOOD, I. J. (1986c). The whole truth. *IMS Bull.* **15** 366–373.
GOOD, I. J. (1986d). The computer-intensive form of a Bayes/non-Bayes compromise. *J. Statist. Comput. Simulation* **26** 132–133.
GOOD, I. J. (1986e). A flexible Bayesian model for comparing two treatments. *J. Statist. Comput. Simulation* **26** 301–305.
GOOD, I. J. (1986f). A pragmatic theory of truth of theories or hypotheses. *J. Statist. Comput. Simulation* **24** 320–321.
GOOD, I. J. (1987a). Causal tendency: a review. A revision of Good (1985d) as an invited paper at the Conference on Probability and Causation at Irvine, California, 1985, July 15–19. In *Causation, Chance, and Credence* (W. Harper and B. Skyrms, eds.) 23–50. Reidel, Dordrecht.
GOOD, I. J. (1987b). A reinstatement, in response to Gillies, of Redhead's argument in support of induction. *Philos. Sci.* **54** 470–472.
GOOD, I. J. (1987c). Subjective probability. In *The New Palgrave: a Dictionary of Economics* (J. Eatwell, M. Milgate and P. Newman, eds.) **4** 537–543. Stockton Press, New York.
GOOD, I. J. (1988a). Scientific method and statistics. In *Encyclopedia of Statistical Sciences* **8** 291–304. Wiley, New York.
GOOD, I. J. (1988b). Surprise index. In *Encyclopedia of Statistical Sciences* **9** 104–109. Wiley, New York.
GOOD, I. J. (1988c). Statistical evidence. In *Encyclopedia of Statistical Sciences* **8** 651–656. Wiley, New York.
GOOD, I. J. and CARD, W. I. (1971). The diagnostic process with special reference to errors. *Methods Information Med.* **10** 176–188.
GOOD, I. J. and CROOK, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69** 711–720.
GOOD, I. J. and CROOK, J. F. (1987). The robustness and sensitivity of the mixed Dirichlet Bayesian test for "independence" in contingency tables. *Ann. Statist.* **15** 670–693.
GOOD, I. J. and DEATON, M. L. (1981). Recent advances in bump-hunting (with discussion). *Computer Science and Statistics: Proc. of the 13th Symposium on the Interface* (William F. Eddy, ed.) 92–104. Springer, New York.
GOOD, I. J. and GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58** 255–277.
GOOD, I. J. and GASKINS, R. A. (1972). Global nonparametric estimation of probability densities. *Virginia J. Sci.* **23** 171–193.
GOOD, I. J. and GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion). *J. Amer. Statist. Assoc.* **75** 42–73.
GOOD, I. J. and MCMICHAEL, A. F. (1984). A pragmatic modification of explicativity for the acceptance of hypotheses. *Philos. Sci.* **51** 120–127.
GOOD, I. J. and TOULMIN, G. H. (1968). Coding theorems and weight of evidence. *J. Inst. Math. Applications* **4** 94–105.
HICKEY, H. (1987). The Challenger tragedy: It exposed TV's failures as well as NASA's. *TV Guide* January 24–30, 2–11.
JEFFREY, R. C. (1965). *The Logic of Decision.* McGraw-Hill, New York.
JEFFREYS, H. (1939). *Theory of Probability.* Clarendon Press, Oxford.
JOHNSON, W. E. (1932). Appendix to "Probability: Deductive and inductive problems" edited by R. B. Braithwaite. *Mind* **41** 421–423.
KEMBLE, E. C. (1942). Is the frequency theory of probability adequate for all scientific purposes? *Amer. J. Phys.* **10** 6–16.
KEYNES, J. M. (1921). *A Treatise on Probability.* Macmillan, New York; St. Martin's Press, 1952.
LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **37** 23–37.
LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.* **72** 869–874.
LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.
LINDLEY, D. V. (1956). On the measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.
LINDLEY, D. V. (1971). The estimation of many parameters (with discussion). In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 435–455. Holt, Rinehart and Winston, Toronto.
LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 1–42.
OGDEN, C. K. (1959). *Bentham's Theory of Fictions.* Routledge and Kegan Paul, London.
PEARSON, E. S. (1955). Statistical concepts in their relation to reality. *J. Roy. Statist. Soc. Ser. B* **17** 204–207.
PEIRCE, C. S. (1878). The probability of induction. *Popular Science Monthly.* Reprinted in *The World of Mathematics* (J. R. Newman, ed.) **2** 1341–1354. Simon and Schuster, New York, 1956.
POISSON, S. D. (1837). *Recherches sur la Probabilité des Jugements.* Bachelier, Paris.
POPPER, K. R. (1959). *The Logic of Scientific Discovery.* Hutchinson, London.
POPPER, K. R. and MILLER, D. (1983). A proof of the impossibility of inductive probability. *Nature* **302** 687–688; **310** 434.
POPPER, K. R. and MILLER, D. W. (1987). Why probabilistic support is not inductive. *Philos. Trans. Roy. Soc. London Ser. A* **321** 569–591.
RAMSEY, F. P. (1926/64). Truth and probability. A 1926 lecture published in *The Foundations of Mathematics and Other Logical Essays* (1950). Routledge and Kegan Paul, London. Reprinted in 1964 in *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds.) 63–92. Wiley, New York. In 2nd ed., 1980, 23–52. Krieger, Huntington, N. Y.
REDHEAD, M. L. G. (1985). On the impossibility of inductive probability. *British J. Philos. Sci.* **36** 185–191.
SAVAGE, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
SPIEGELHALTER, D. J. and KNILL-JONES, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology (with discussion). *J. Roy. Statist. Soc. Ser. A.* **147** 1–34.
STIGLER, S. M. (1986). *The History of Statistics: the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard Univ. Press, Cambridge, Mass.
SUPPES, P. (1970). *A Probabilistic Theory of Causality.* North-Holland, Amsterdam.
TURING, A. M. (1941). Private communication.
WAHBA, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering* (J. Van Ryzin, ed.) 423–458. Academic, New York.
WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.
WRINCH, D. and JEFFREYS, H. (1921). On certain fundamental principles of scientific inquiry. *Philos. Mag. Ser.* **6** 369–390.