Statistical Science
2004, Vol. 19, No. 1, 44–57
DOI 10.1214/088342304000000125

Probability, Causality and the Empirical World: A Bayes–de Finetti–Popper–Borel Synthesis

A. P. Dawid

Abstract. This article expounds a philosophical approach to Probability and Causality: a synthesis of the personalist Bayesian views of de Finetti and Popper's falsificationist programme. A falsification method for probabilistic or causal theories, based on "Borel criteria," is described. It is argued that this minimalist approach, free of any distracting metaphysical inputs, provides the essential support required for the conduct and advance of Science.

Key words and phrases: Borel criterion, calibration, falsification, Jeffreys's law.

1. INTRODUCTION

The neo-Bayesian revival of the 20th century was stimulated by the contributions of such workers as Ramsey (1926), de Finetti (1974–1975), Savage (1954) and Lindley (1965), with their strong emphasis on *personal probability* (more often, but less appropriately, called *subjective probability*)—an interpretation that is, however, at best only implicit in the original memoir of Bayes (1763). Conversely, there is relatively little attention paid by personalist Bayesian statisticians (although more by Bayesian philosophers: see, e.g., Howson and Urbach, 1993) to what was arguably Bayes's principal purpose: to provide a machinery for drawing *causal conclusions*. Thus if C is a possible cause of an observed effect E, Bayes's theorem provides exactly the requisite logic to assess $P(C|E)$, the believability of cause C in the light of the observation, in terms of the more basic ingredients $P(E|C)$, the probability of obtaining the effect E when cause C is in fact operating, and $P(C)$, the prior believability of C.

The personalist Bayesian viewpoint remains controversial, even among Bayesians. The principal objection has been that Science is, or should be, the search for objective facts and laws, together with an associated view that the instruments and arguments used in that search should likewise have an objective status, independent of any personal views of the Scientist. I myself have been enormously influenced by the logical cohesion and persuasiveness of the uncompromisingly personalist approach to Probability set out by de Finetti. At the same time I do take seriously the criticism that, if this theory is to be more than a branch of Psychology, some further link with external reality is needed.

In this article I describe my attempt to relate personal probabilities to the external world. In fact the scope of this program is wider, since the personalist—or any other—interpretation of probability is not essential to it. I consider general scientific theories of the world, expressed mathematically in a way that involves probabilistic terms (supposed to satisfy Kolmogorov's axioms), and ask what it means for such a theory to do a good or bad job of describing the world. The suggested answer, based on Borel's "single law of chance," is indirect and subtle, but I believe it serves the desired purpose well.

I also consider here the empirical content and meaning of *causal* probabilistic models. These have recently attracted considerable interest in the general statistical community (Rubin, 1978; Robins, 1989; Pearl, 2000; Shafer, 1996). In Dawid (2000) I pointed out some worrying metaphysical aspects of currently popular approaches and outlined an alternative decision-theoretic approach which avoids these. This is a straightforward extension of regular probabilistic modelling and is amenable to analysis by the identical methods and tools. Consequently, we can apply the philosophical

A. P. Dawid is Professor, University College London, Gower Street, London WC1E 6BT, UK (e-mail: dawid@stats.ucl.ac.uk).

understandings developed for probabilistic models to understand causal models also.

I have been developing the principal ideas described here over many years. For the relationship between probability modelling and reality, see Dawid (1982b, 1984c, 1985a–c), Seillier-Moiseiwitsch and Dawid (1993), Dawid (1997) and Dawid and Vovk (1999). For exchangeability and its extensions, see Dawid (1977, 1982a, 1984b), Consonni and Dawid (1985) and Dawid (1985c, 1986a, 1988). For causality, see Dawid (1979, 1984a, 2000, 2002, 2003). The current article attempts to provide a "position statement," locating these scattered contributions within a (reasonably) coherent overall philosophy.

1.1 Preview

Section 2 briefly describes the two philosophies, due respectively to de Finetti and Popper, that have inspired my own approach to understanding probabilistic models of the world. Although these thinkers would generally be regarded as having very different, even mutually inconsistent, viewpoints, I believe that a synthesis of their principal points of emphasis is both possible and valuable.

In Section 3, I review and exemplify various ideas that have been propounded, from differing philosophical standpoints, on the meaning of probability and its real-world interpretation. In particular, Section 3.3.4 describes an indirect approach achieving the desired de Finetti–Popper synthesis. Section 4 then uses this as a basis for a general method, based on "probability calibration," for testing probabilistic models through appropriate comparisons of theoretical probabilities and observed frequencies. The argument is extended to causal modelling in Section 5. Section 6 contains some concluding remarks.

2. TWO PHILOSOPHIES

2.1 de Finetti

de Finetti's philosophical approach might be described as "Machian," or "extreme positivist." It ascribes meaning only to those events and quantities that are *observable* in the world and provides various measuring instruments (e.g., gambling scenarios, proper scoring rules, decision problems) that can be used to quantify Your[1] uncertainty about any such unknown quantity. In restricting the scope of uncertainty judgments to genuine observables, de Finetti can himself be regarded as following in the footsteps of Bayes (1763), who derived the uniform prior distribution for an unobservable binomial probability parameter from the more basic judgment of a discrete uniform predictive distribution for the number of successes in the next n trials.

de Finetti showed that *coherence*, a simple economic behavioral criterion—essentially, that You ought to avoid a combination of decisions that is guaranteed to lead to loss—is all that is needed to ensure that Your uncertainty can be represented and manipulated using the mathematical theory of probability, including, as an important special case, Bayes's theorem.

2.1.1 *Exchangeability.* Perhaps the greatest and most original success of de Finetti's methodological program is his theory of *exchangeability* (de Finetti, 1937). When considering a sequence of coin-tosses, for example, de Finetti does not assume—as would typically be done automatically and uncritically—that these must have the probabilistic structure of Bernoulli trials. Instead, he attempts to understand when and why this Bernoulli model might be reasonable. In accordance with his positivist position, he starts by focusing attention directly on Your personal joint probability distribution for the potentially infinite sequence of *outcomes* $(X_1, X_2, \ldots)$ of the tosses—this distribution being numerically fully determined (and so, in particular, having no "unknown parameters"). Exchangeability holds when this joint distribution is symmetric, in the sense that Your uncertainty would not be changed even if the tosses were first to be relabelled in some fixed but arbitrary way (so that, e.g., X_1 now refers to toss 5, X_2 to toss 21, X_3 to toss 1, etc.). In many applied contexts You would be willing to regard this as an extremely weak and reasonable condition to impose on Your personal joint distribution, at least to an acceptable approximation. de Finetti's famous representation theorem now implies that, assuming *only* exchangeability, we can deduce that Your joint distribution is exactly the same *as if* You believed in a model of Bernoulli trials, governed by some unknown parameter p, and had personal uncertainty about p (expressed by some probability distribution on $[0, 1]$). In particular, You would give probability 1 to the existence of a limiting relative frequency of H's in the sequence of tosses, and could take this limit as the definition of the "parameter" p. Because it can examine frequency conceptions of Probability from an external standpoint, the theory of personal probability is able to

[1] Following de Finetti, we denote by "You" the subject whose uncertainty and personal probabilities are under consideration.

cast new light on these—an understanding that is simply unavailable to a frequentist, whose very conception of probability is already based on ideas of frequency. Even more important, from this external standpoint these frequency interpretations are seen to be relevant only in very special setups, rather than being fundamental: for example, there is no difficulty in principle to extending the ideas and mathematics of exchangeability to two-dimensional, or still more complicated, arrays of variables (Dawid 1982a, 1985c).

2.2 Popper

The scepticism toward the personal Bayesian approach among many statisticians might be countered by exhibiting some way of confronting Your psychological uncertainty assessments with what actually happens in the world. One such approach (see Dawid, 1986b) can be based on "proper scoring rules"—quantative measures of the success of probability evaluations in the light of observed outcomes, which can be used to compare rival evaluations. Here, however, we shall be concerned with absolute, rather than relative, assessments of empirical adequacy.

The falsificationist approach of Karl Popper (1959) takes this issue seriously. Popper classifies theories as either "physical" or "metaphysical." A physical theory makes predictions about the real world that can be tested by observation. If such a prediction fails, the theory is discredited and must be discarded or modified. However, passing such a test does not in itself render the theory "proven" or "true" in any sense—indeed, from a thoroughgoing falsificationist standpoint (perhaps even more thoroughgoing than Popper himself would have accepted), we can dispense with such concepts altogether. It may well be the fate of any theory we can ever devise to be eventually falsified; we can only hope that our currently unfalsified theory or theories will prove useful for understanding and predicting the world, during their limited lifespan. Popper's approach takes seriously Hume's argument that there can be no noncircular argument justifying induction, but, rather than throwing in the towel, decides to get on with the job and hope for the best. It is better described as a methodological program rather than a philosophy. It is entirely agnostic about such deep matters as the "truth" of intellectual theories. Indeed, one does not have to subscribe to the view that falsification is all there is to science to find it a fruitful approach: it is in essence a minimalist program, and any understandings that can be reached by this means, without any further philosophical inputs or assumptions, will have a correspondingly high degree of robustness.

A theory that makes purportedly meaningful assertions that cannot be falsified by any observation is "metaphysical."[2] Whatever other valuable properties such a theory may have, it would not, in Popper's view, qualify as a *scientific* theory. While Popper did not disdain metaphysics, he wished to demarcate the boundary between metaphysics and science, claiming falsifiability as the defining feature of the latter, and aiming to remove metaphysical concepts from scientific theories as much as possible.[3]

Although Popper disagreed vehemently with inductivist approaches to science such as that of de Finetti, at a deep level it seems that both the positivist de Finetti and the "negativist" Popper shared much the same empiricist view of the nature and importance of the relation between theory and reality.

3. THEORY AND REALITY

While there are many subtle philosophical issues involved in putting Popper's falsificationist program into effect, I find the overall approach it embodies just as

[2] We can distinguish various degrees of falsification of a theoretical assertion, and correspondingly varying degrees of (meta)physicality. Popper himself talks of "basic statements," themselves expressible within the theory, whose truth would be inconsistent with that theory—but leaves somewhat vague the empirical status of such statements. When such a basic statement can be interpreted as describing a currently possible, or at least conceivable, measurement or set of measurements, we might call it *physical*. A statement that is currently unfalsifiable, but might become so by extensions to our theory and/or experimental abilities, might be termed *potentially physical*. A statement that could not, on purely logical grounds, be falsifiable, even with such extensions, is *fundamentally metaphysical*. The boundaries are not always clear: for example, are statements about propensities (see Section 3.3.2) potentially physical or fundamentally metaphysical (I opt for the latter, but could understand the other viewpoint)? However, I consider some of the statements made in potential response theories of causality (see Section 5.1) as, indubitably, fundamentally metaphysical.

[3] It is interesting that Popper himself pays great attention to such metaphysical ideas as realism (which he accepts) and idealism and instrumentalism (which he rejects, but which have been guiding spirits behind my own program). Thus he was a firm believer in the meaningfulness of the concept of the *truth* of a scientific theory, and even seems to have believed that some current scientific theories were indeed true (although we could never know this for sure). It seems to me that such considerations are simply irrelevant to his falsificationist methodological program, which can be clarified and strengthened by applying to it the the same exhortation to manage without metaphysical concepts that it itself imposes on specific scientific theories.

compelling as de Finetti's arguments for coherence of personal probabilities. Over many years I have worked toward forming a synthesis of these two approaches, by developing falsification criteria to assess the empirical performance of Your personal probability judgments. Such a criterion can then be used more generally, to assess (always purely tentatively) the empirical worth of any proposed probability model, irrespective of its origins, be they personalist or otherwise.

3.1 Two Universes

For sensible discussion of these issues, I regard it as of vital importance to distinguish, carefully and constantly, between two very different universes, which I will term "intellectual" and "physical."[4]

Any kind of scientific, mathematical or logical theory is a purely intellectual construct. It will typically involve a variety of symbols and concepts, together with rules for manipulating them. The physical universe, on the other hand, just does its own thing, entirely ignorant of, and careless of, any of our intellectual theories. It manifests itself to us by means of observations. I consider that failure to appreciate which of these universes is appropriate to some topic or term of discourse is a common cause of confusion, with potentially seriously misleading consequences.

A specifically *scientific* theory—unlike, for example, a purely mathematical one—further purports to "say something about" the physical universe. However, in order for it to be able to do this there must first be established some clearly understood *link* between the two universes, intellectual and physical. Such a link will involve explicit or implicit *rules of interpretation*. For example, we might agree to interpret the symbol X in a certain application of a certain theory as representing the distance between the earth and the sun, as determined by a suitable measurement protocol.

3.2 Deterministic Theories

With important exceptions such as quantum mechanics and Mendelian genetics, most modern scientific theories are deterministic. They involve terms whose theoretical relationships can be understood, using the relevant rules of interpretation, as relationships that should be satisfied by certain measurements in the physical universe. If we can make these measurements, we can look to see whether or not the asserted relationships hold. Any failure to do so would falsify the model.

3.3 Probabilistic Theories

There are serious difficulties, however, in applying the above naïve falsificationist program directly to a probabilistic theory. Thus suppose that we are to toss a given coin repeatedly, and observe the outcome, H or T. Consider the theory wherein we represent the outcome of the ith toss of this coin by X_i, further representing outcome H by 1 and T by 0; and model the (X_i) by means of the joint probability distribution P under which they are independent, with $P(X_i = 1) = 0.5$. We call this the "fair coin model."

How could we falsify such a probabilistic theory? As a basis for this, we would first need to supply appropriate rules of interpretation, linking the probabilistic ingredients of the model to the physical universe. We now consider a number of ways in which this task might be approached.

3.3.1 *Personal probability.* If the fair coin model P is interpreted as merely describing Your personal beliefs, it could be tested without even tossing the penny, by observing Your behavior, for example in various gambling scenarios. However, this purely psychological falsifiability clearly misses the main point: that the theory is intended to describe the external physical universe, rather than (or at least, in addition to) Your internal psychological state.

It is sometimes suggested that the theory of exchangeability, or its variations, supplies a way of linking psychological and physical probabilities. Thus in the case described in Section 2.1.1 of exchangeable personal opinions about a sequence of coin tosses, de Finetti's theorem appears to prove, mathematically, the existence of a limiting relative frequency p conditional on which the coin behaves according to a Bernoulli model with probability parameter p—and this quantity p would then seem to be a suitable candidate for the label "physical probability." The fair coin model can be regarded as obtained by combining exchangeability with the additional requirement $p = 0.5$, and this latter requirement (together with other implications of the Bernoulli model) could in principle

[4] The latter ("physical") corresponds roughly to Popper's World 1 (Popper, 1982); the former ("intellectual") could refer equally to his World 2 (comprising subjective thought and experience) or World 3 (comprising "objective thought"). Our distinction might appear to presuppose a dualist philosophy, but this is inessential: the "physical universe" could itself be given an idealist, even solipsistic, phenomenological interpretation, in terms of Your perceptions of sense impressions, thus reducing World 1 to World 2 (we here gloss over the difficulties involved in abstracting from sense perceptions to supposed "real-world" quantities).

be tested empirically (see Sections 3.3.3 and 3.3.4). However, the basic assumption of exchangeability is itself physically untestable. Furthermore the qualification "with probability 1" governing the existence of p in de Finetti's theorem is no mere technicality, but refers to Your psychological belief state, which could be quite out of touch with reality. All that de Finetti's theorem in fact shows is that, if You are coherent and Your opinions about the coin tosses exhibit exchangeability, then this commits You to *believe* in (i.e., attach probability 1 to) the existence of p. But this quantity still inhabits the intellectual universe: Your belief can give no guarantee of the *actual* existence of a counterpart of p in the physical universe. Indeed, as we shall see in Section 3.3.4, it is just this important distinction, between firmly held intellectual beliefs and physical reality, that is fundamental to our interpretation of probability models. Failure to appreciate it is liable to lead to serious misinterpretations (Kalai and Lehrer, 1993; Miller and Sanchirico, 1999).

3.3.2 *Propensity.* In Popper's *propensity theory* of probability (Popper, 1983), which superseded his initial frequency interpretation (Popper, 1959), there is supposed to exist in the physical universe, and quite independent of our intellectual theorizing, a "propensity" for a particular toss (say the ith) of the penny to result in H;[5] this is taken as the intended external referent of the theoretical term $p_i := P(X_i = 1)$ of the fair coin model. This model would then necessarily be false if the physical propensity to obtain H on toss 1 were not in fact 0.5. Likewise, we might contemplate the physical existence of a *joint propensity* to obtain the result H on both tosses 1 and 2: the model would equally be false if this joint propensity were not 0.25.

Unfortunately, it is difficult to see how to translate this understanding of *falsity* into a workable *falsification criterion*—since we cannot construct or even conceive[6] of any measurement protocols to determine directly any physical propensity, such as that of obtaining H on toss 1 of the penny. It seems to me that Popper would have to classify his own propensity theory as metaphysical.

3.3.3 *Frequency.* A frequentist such as von Mises (1939) would deny the physical existence of unique-case propensities, while admitting the physical existence of a "generic" probability for the given coin to fall H in tosses of this kind. This could be defined directly in terms of more basic physical observables such as the limiting proportion of H's obtained on tossing the coin repeatedly. From this point of view, the fair coin model would be falsified if this limit fails to exist and to equal 0.5.

Refinements of this idea can be developed to address, additionally, the important *independence* property embodied in the fair coin model. Thus von Mises requires that the sequence of observed outcomes satisfy his "randomness postulate." Informally, this means that the property of yielding limiting relative frequency 0.5 should obtain, not just for the complete sequence of outcomes, but also for subsequences, chosen by "place-selection rules" whereby the decision whether or not to include any particular toss in the subsequence can depend on the outcomes of previous tosses. Although not mathematically watertight in its original formulation, this conception can be made rigorous (see, e.g., Dawid, 1985a, Section 2).

Sophisticated and imaginative though such approaches are, it could be argued that they do not really address the fair coin Bernoulli trials model directly, but subject it to considerable reinterpretation and distortion in their attempts to link it to the physical universe.

3.3.4 *Borel criteria.* My own favored approach avoids completely any need to assume the physical existence of probabilities, however interpreted. I regard "probability" as a *purely theoretical* term, inhabiting the intellectual universe and without any direct physical counterpart. We might interpret such a theoretical probability as a personal degree of belief, or as a propensity, but any such interpretation is irrelevant to our purposes. Rather, we should regard probabilities as entering our scientific theories as *instrumental* terms, the link between theoretical probabilities and the physical universe being indirect. This approach to interpreting probabilistic models avoids many potential philosophical pitfalls. In particular, by treating probabilities as purely theoretical terms with only indirect implications for the behavior of observables, it is able to eschew deep but ultimately irrelevant and distracting philosophical inquiry into the "true nature of Probability."

We take as our falsifiability criterion (more precisely, a family of criteria) a version of a principle that, in one form or another, has been propounded by Bernoulli (1713), Cournot (1843) (see Shafer and Vovk, 2001) and Borel (1943), who called it "the Single Law of Chance." We regard a probabilistic theory as falsified

[5] More precisely, Popper considers a propensity as associated with a particular "experimental arrangement," detailing the conditions under which the outcome is generated.

[6] At any rate, I cannot—see footnote 2.

if it assigns probability unity to some prespecified theoretical event A, and observation shows that the physical counterpart of the event A is in fact false. We shall call such an event A a "Borel criterion." In the minimalist[7] version of this approach, only extreme (i.e., 0 or 1) probabilities in a theory are regarded as having any meaningful direct external referent, and only such extreme probability statements can act as potential falsifiers of the theory.

For example, for our fair coin model it is a mathematical theorem that the *theoretical event*

$$n^{-1}\sum_{i=1}^{n} X_i \to 0.5$$

has P-probability 1. Hence, as a model of the physical universe, P could be regarded as falsified if, on observation, the corresponding physical property,

> the limiting relative frequency of H in the sequence of coin-tosses exists and equals 0.5,

is found to fail. In this simple case, this conclusion happens to agree with that of Section 3.3.3. The same reasoning can be used to justify the extensions mentioned there to test independence, since these all refer to properties that are in fact assigned theoretical probability 1 by the Bernoulli model P. Borel criteria, however, are much more general: in particular, they need in no way rely on any kind of "repeated trials" setup.

It is worth remarking that no Borel criterion could ever be able to distinguish between two probabilistic models, P and Q, that are mutually absolutely continuous ($P \approx Q$)—that is, which agree as to which events are assigned probability 0 (although not necessarily on other probabilities), and thus both pass or fail any Borel falsifiability test together. This implies that, with this approach, we can never home in on a unique "true model" by eliminating all others. This is a strong, and perhaps *prima facie* disturbing, instance of the extreme falsificationist position that there is just no such thing as "the true model" for (some aspect of) the physical world.

For example, if, under Q, the (X_i) are taken as independent, with $Q(X_i = 1) = q_i$, where $0 < q_i < 1$ and $\sum_{i=1}^{\infty}(q_i - 0.5)^2 < \infty$, while P is the fair coin model, then it follows mathematically that $P \approx Q$, so that no Borel criterion will ever be able to distinguish between P and Q.[8] In particular, no specific nonextreme assignment of a value q_i to the probability of obtaining H on toss i can ever be falsified. Typically only a complete probabilistic theory, involving an infinite collection of probability assignments, can be falsified using a Borel criterion.

Although the inability to discriminate between two models P and Q when $P \approx Q$ might seem to be a fundamental problem with this approach, it can be argued that, from a pragmatic point of view, it is of no importance. For it can be shown (Blackwell and Dubins, 1962) that when this absolute continuity holds, as the current time $N \to \infty$, the conditional distribution of (the remainder of) the full sequence $\mathbf{X}$ given the currently available information $(X_1, \ldots, X_N)$ will be, in a strong sense, asymptotically the same, whether it is calculated under P or under Q—this property holding almost surely under both P and Q. Consequently, given enough data, both P and Q will make essentially identical predictions, and it simply will not matter which one we use. This reassuring property is an instance of what I have termed "Jeffreys's law" (Dawid, 1984c): distinct theories that cannot eventually be distinguished empirically do not in fact need to be so distinguished.

The above approach to the testing of probabilistic models, using Borel criteria, takes very seriously de Finetti's motto "Probability does not exist" (de Finetti, 1974–1975): that is to say, it denies—or at any rate, has no use for—the idea that there is a property of the physical world that is in direct correspondence to the theoretical concept of probability. This necessitates indirect application of Popper's hypothetico-deductive approach. We suppose that, somehow or other,[9] we have come up with one or more theories, involving probabilities (perhaps, but not necessarily, having a personalist interpretation) as theoretical terms. These are attached as attributes to other

[7] Unlike personalist interpretations, or that of Popper (1983), who regards Borel criteria as providing a "bridge" between empirical observations and a probabilistic theory interpreted as making assertions about propensities, considered as meaningful (though unobservable) real-world quantities.

[8] In fact we can even allow dependence among the (X_i) under Q. The general condition for $P \approx Q$ is $\sum_{i=1}^{\infty}(q_i - 0.5)^2 < \infty$ (Q-almost surely), where now $q_i := Q(X_i = 1 | X_1, \ldots, X_{i-1})$.

[9] I do not propose that the falsificationist methodological program laid out here is in any way relevant to the important, but separate, issues of *hypothesis generation* and *hypothesis interpretation*. My own preference is for hypotheses that can be interpreted as tentative personalist views of the world, but this is inessential. I would not deny that a variety of interpretations (e.g., based on a realist view of Nature, ascribing truth or meaning to terms that are later regarded purely instrumentally for falsificationist purposes) can be fruitful in suggesting interesting models and hypotheses.

terms that name well-specified physically observable quantities and events. For such a model, we construct an appropriate Borel "test event" that is assigned probability 1 by that model; and we reject, or modify, any model that fails its test.[10] Any remaining models can then be used, albeit tentatively, for prediction, or other such purposes—although these, too, may also eventually be rejected. We have no need to believe in the existence of a true model, describing the physical world perfectly (which is to say, never being rejected by our falsification criterion); and even were this to be the case, such a "true" model would typically not be unique! However, Jeffreys's law suggests that all models that survive sufficiently intensive testing should be essentially equivalent for predictive purposes.[11]

3.4 An Example

To elaborate further the differences between the above viewpoints, we consider a slightly more complicated problem (Dawid, 1985c). Suppose that it is possible to sample coins from the potentially infinite production of a mint, and to toss any one of these as often as desired. Let X_{ij}, coded 0 or 1 as before, represent the outcome of toss j of coin i.

3.4.1 *Personal probability.* As a personalist, You might structure Your personal opinions having regard to Your perceptions of symmetries in the problem. Thus, in an extension of the idea of exchangeability (see Section 2.1.1), You might consider using a joint distribution P for all the (X_{ij}) that would be unchanged if someone were first to permute the ordering of the coins, or to permute the ordering of the tosses of any fixed coin. These personal judgments would be reflected in Your indifference between bets, whether placed before or after the application of such permutations. This observable property could be used to test the validity of P as a model of Your internal *psychological* state, but not its external *physical* validity.

Now it can be shown (Dawid, 1985c) that the above symmetry assumptions alone (in particular, without imposing any further assumptions, such as independence) are sufficient to justify the *hierarchical model* M described by the following three stages:

1. A distribution Π over $[0, 1]$ is determined (possibly randomly, by some given process).
2. Conditional on Π, quantities p_i $(i = 1, 2, \ldots)$ in $[0, 1]$ are generated independently from the distribution Π.
3. Conditional on Π and $\mathbf{p} := (p_1, p_2, \ldots)$, the X_{ij} are independent with

$$\text{prob}(X_{ij} = 1 | \Pi, \mathbf{p}) = p_i.$$

The resulting marginal joint probability distribution of the (X_{ij}) will then clearly have the desired symmetries; and in fact any such symmetric distribution can be regarded as having been generated in this way. From the personalist point of view of Section 3.3.1, the model M is no more nor less than a probabilistic representation of Your perceived symmetries in the collection (X_{ij})—in particular, the quantities Π and (p_i) can be regarded as purely "instrumental fictions" for structuring Your joint uncertainty for the observables.

An extension of the above symmetry modelling approach to apply to more general analysis of variance type problems can be found in Dawid (1988).

3.4.2 *Propensity.* How might we interpret the above model M from the propensity viewpoint of Section 3.3.2?

We might first think of the theoretical term p_1 in M as referring to the physical propensity that the first coin will fall H on the first toss—although this interpretation is somewhat muddied by the fact that the (p_i) are themselves being modelled as random. We could perhaps go on to regard Π as referring to a "second-order propensity distribution," again living in the physical universe, that governs the assignment of a value to the propensity, represented by p_1, that the first coin will fall H on the first toss.

Now consider the collection of all *first* tosses of all the coins. Under the model M the corresponding quantities (X_{i1}) $(i = 1, 2, \ldots)$ behave as Bernoulli trials with probability parameter

$$p^* := \int_0^1 u \, d\Pi(u).$$

The theoretical quantity p^* appears to be just as strong a candidate as the quantity p_1 to act as the theoretical counterpart of the physical propensity that the first toss of the first coin will result in H. However, in general $p^* \neq p_1$, whereas the above propensity, being assumed to exist in the physical universe, must necessarily have a unique value. Since, as noted earlier, there appear to be no instruments for measuring propensities, it is not

[10] We admit, as a potentially serious conceptual and practical difficulty, that typically to conduct such a definitive test will require an infinite number of observations. Within our framework there seems to be no way of avoiding this problem (Dawid, 1985a).

[11] This solves, or at any rate sidesteps, Goodman's "new riddle of induction" (Goodman, 1954, Chapter III) since sufficiently intensive testing over time would serve to distinguish between the nonequivalent theories "all emeralds are green" and "all emeralds are grue."

clear how this ambiguity in propensity theory could be resolved.[12]

3.4.3 *Frequency.* The frequentist position of Section 3.3.3 is strongly tied to models involving repeated trials under identical conditions. In level 3 of model M we do have such a situation for each individual coin i. Then we can regard p_i as referring to the limiting relative frequency of H in tosses of coin i.

Since M does not assign a specific value to p_1, we cannot use this as the basis of a falsification test; however, we can apply the extensions mentioned in Section 3.3.3 to see whether we can falsify the implicit assumption of independence between tosses of the coin.

At level 2 of model M, we can consider the collection (p_i) of all the observed limiting relative frequencies, across all coins, and look to see whether this is consistent with our model assumption that these arise by repeated sampling from the distribution Π. (If Π is not itself fully specified by level 1, we can at least examine the "independent and identically distributed" assumption.)

Since there is no replication at level 1 of our model, there is no possibility of falsifying it at that level by any frequentist criterion.

3.4.4 *Borel criteria.* The procedure described in Section 3.4.3 seems to be an essentially straightforward extension of the frequentist approach to deal with this more complex "nested" structure. However, this frequentist logic cannot be naturally extended to more complex structures, such as a cross-classified layout, which do not involve any obvious embedded "repeated trials." By contrast, the approach of Section 3.3.4 does not rely on any repeated trials structure and can be applied equally well to such more general structures.

For a wide variety of problems, the theory of *extreme-point modelling* (Lauritzen, 1988) can be used to pass from broad-brush personal judgments (e.g., Dawid, 1982a, in a generalization of the theory of exchangeability, from perceptions of invariance under a relevant transformation group; or Diaconis and Freedman, 1980, from considerations of sufficiency) to corresponding statistical models. This theory can be used to justify the hierarchical model M of Section 3.4.1 (with Π regarded as a nonrandom parameter), as well as models for cases such as a cross-classified layout (Aldous, 1981; Lauritzen, 2003) and still more complex extensions.

Once an appropriate joint probability model has been constructed, we can analyze it mathematically to identify suitable probability-1 events to use as Borel criteria for falsification tests. However, this analysis may be nontrivial, and the choice between different Borel criteria nonobvious.

4. CALIBRATION AND EXTENSIONS

How, in general, might we select the test event A to use as a Borel criterion? In Dawid (1985a) I suggested that a suitable test could be based on the property of *computable calibration*. This applies directly when quantities are observed in sequence, although it can also be applied much more generally by first ordering the quantities somehow—the exact ordering being asymptotically unimportant.

Suppose that P is an arbitrary joint distribution for a potentially infinite sequence $\mathbf{X} := (X_1, X_2, \ldots)$ of quantities, for simplicity here supposed binary, and that we wish to test the suitability of P as a model, in the light of a sequence $\mathbf{x} := (x_1, x_2, \ldots)$ of empirical outcomes. We introduce the random quantity $\Pi_i := P(X_i = 1|X_1, \ldots, X_{i-1})$, a function of the quantities $(X_1, \ldots, X_{i-1})$ which also depends on the model P, and its observed value $\pi_i := P(X_i = 1|X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$, which depends both on the observed values $(x_1, \ldots, x_{i-1})$ and on the model P. Without needing to impose any further conditions, such as independence, it can be shown (Dawid, 1985a) that P assigns probability 1 to the event

$$C_0\text{: } \text{“}\overline{X}_N - \overline{\Pi}_N \to 0 \quad \text{as } N \to \infty,\text{”}$$

where $\overline{X}_N := (1/N)\sum_{i=1}^{N} X_i$, etc. Using this event to construct our Borel criterion, we could regard the model P as falsified by the empirical observations if, in fact, $\overline{x}_N - \overline{\pi}_N \not\to 0$. This property, "overall calibration," requires that, in the long run, the average of all the actually emitted sequential probability forecasts π_i should agree with the relative frequency of outcome 1 among the observed values (x_i) of the (X_i).

The above test is relatively crude and undiscriminating: for example, it would not reject the fair coin model, with its constant forecast sequence $(0.5, 0.5, 0.5, \ldots)$, when the actual outcome sequence alternated as $(0, 1, 0, \ldots)$—looking very far from random.

To construct more refined tests, for each i let Δ_i be a function of $(X_1, \ldots, X_{i-1})$ taking values in $\{0, 1\}$.

[12] One attempted solution might be to say that the two propensity values refer to different experimental conditions: but if the experiment involves selecting one coin at random and tossing it once, it is not clear how its conditions change merely by changing the sequence within which we consider this toss embedded.

We can collect together these functions as a single function (place-selection rule) Δ defined on all finite strings, with $\Delta(x_1, \ldots, x_{i-1}) := \Delta_i(x_1, \ldots, x_{i-1})$. For any infinite sequence $\mathbf{x} = (x_1, x_2, \ldots)$, Δ determines a subsequence $\mathbf{x}_\Delta := (x_i : \delta_i = 1)$, where $\delta_i := \Delta(x_1, \ldots, x_{i-1})$—picking out terms in a way that can depend (in an essentially arbitrary fashion) on previous outcomes.

Define, for any N,

$$N_\Delta := \sum_{i=1}^{N} \Delta_i, \quad \overline{X}_\Delta := \frac{\sum_{i=1}^{N} \Delta_i X_i}{N_\Delta},$$

$$\overline{\Pi}_\Delta := \frac{\sum_{i=1}^{N} \Delta_i \Pi_i}{N_\Delta}.$$

Thus the variables N_Δ, $\overline{X}_\Delta$ and $\overline{\Pi}_\Delta$ are respectively the total count, the relative frequency and the average probability forecast, up to real time N, in the subsequence $\mathbf{X}_\Delta$. We denote the realized values of N_Δ etc., for an observed data-sequence $\mathbf{x} = (x_i)$, by n_Δ etc.

Now consider the event

$$C_\Delta\text{: ``}N_\Delta \to \infty \quad \Longrightarrow \quad \overline{X}_\Delta - \overline{\Pi}_\Delta \to 0\text{,''}$$

which essentially asserts the limiting equality of the average forecast and average outcome over this subsequence (whenever it is infinite). It can be shown that, for any place-selection rule Δ, P assigns probability 1 to the event C_Δ; so again, we could regard P as falsified if, for the empirically realized data-sequence $\mathbf{x}$, $n_\Delta \to \infty$ while $\overline{x}_\Delta - \overline{\pi}_\Delta \not\to 0$.

The "P-almost sure" property is clearly retained if we require such subsequence calibration for every member of some countable collection of place-selection rules Δ. The most incisive such criterion, *computable calibration*, uses as its Borel test criterion the event $C^* := \bigcap_{\Delta \in \mathcal{D}} C_\Delta$, where $\mathcal{D}$ is the countable collection of place selection rules such that Δ is a computable function on its domain (of all finite strings). Then we can regard the probability model P as falsified, in the light of a string $\mathbf{x}$ of observations, if the subsequence calibration property C_Δ fails to hold for some computable place-selection rule Δ applied to $\mathbf{x}$.

The criterion of computable calibration can be regarded as an extension of von Mises's frequentist approach (Dawid, 1985a). Applied to the Bernoulli trials model, it is equivalent to testing the appropriateness of that model by checking whether or not the outcome sequence satisfies the "randomness postulate" (see Section 3.3.3). However, the general calibration criterion is not restricted to Bernoulli trials: it applies to probability models of arbitrary structure and complexity, with no further assumptions such as independence. The only restriction is that we must first order the variables.

4.1 Gambling Systems

An alternative criterion that is still more incisive, and perhaps even more natural, than calibration can be based on the "martingale" approach of Ville (1939). Consider an adversary A, who is trying to discredit Your probability model P. At time i, when you have both observed $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$, You would regard $\pi_i = P(X_i = 1 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$ as a "fair exchange price" for the still-to-be-observed quantity X_i. That is, You would regard as fair a gamble that transfers an amount $c_i(X_i - \pi_i)$ from You to A, whatever the size c_i, be it positive or negative, of the gamble. Moreover, this attitude should not be affected even if A were to adjust the sizes (c_i) of his bets in the light of previous outcomes, taking $c_i = \sigma_i(x_1, \ldots, x_{i-1})$ for some *strategy* $\sigma = (\sigma_1, \sigma_2, \ldots)$ (a real-valued function on finite strings). For any such strategy, You should be happy to accept A's sequence of bets.

Assume that A starts with capital $K_0 = 1$ and operates a strategy σ; then A's capital just after time N is $K_N := 1 + \sum_{i=1}^{N} \sigma(X_1, \ldots, X_{i-1})(X_i - \Pi_i)$. It can be shown that, so long as σ is constructed so that $K_i \geq 0$ for any possible outcome sequence (A is not allowed to risk losing more than he possesses), the distribution P assigns probability 1 to the event that K_N will remain bounded above as $N \to \infty$: that is, if You only take on gambles You regard as fair, You do not expect that you will lose Your whole fortune. Taking this property as our Borel criterion, we can thus regard the distribution P as empirically discredited, in the light of a specific observed sequence $\mathbf{x}$ of outcomes, if $\sum_{i=1}^{N} c_i(x_i - \pi_i) \to \infty$, that is (without ever risking going negative), the adversary's betting strategy succeeds in wiping You out. Once again, we can extend this by allowing some such computable adversary to succeed against You.

This martingale criterion is in some sense the strongest that can be applied, in that when it is satisfied so too are other Borel criteria such as computable calibration. A form of the criterion (without, however, invoking computability) has been taken as the "fundamental interpretive principle" underlying the game-theoretical approach to probability theory expounded by Shafer and Vovk (2001). This approach can be used to derive many familiar limit theorems (such as the law

of the iterated logarithm, Dawid and Vovk, 1999), now guaranteed to hold, not merely with probability 1, but for every individual data-sequence for which the underlying probability model is not falsified (by allowing some adversary to become infinitely rich).

4.2 Empirical Probability

The calibration criterion is thoroughly investigated in Dawid (1985a). In particular it is shown that, if P and Q are two computable joint probability distributions for the sequence $\mathbf{X}$, and both P and Q pass the computable calibration test for a certain empirical data-sequence $\mathbf{x}$, then we must have essential agreement of the forecasts (π_i) and (κ_i) respectively produced by P and Q for that data-sequence: $\pi_i - \kappa_i \to 0$.[13] This is another variant of Jeffreys's law.[14] These asymptotically unique calibrated forecast probabilities might then be considered as having some objective existence in the physical universe. However, it can be shown (Dawid, 1985b) that there is no computable procedure to discover the values of these "empirical probabilities" for any given data-sequence. And it is possible that *no* sequence of forecasts will be computably calibrated for a given data-sequence (Schervish, 1985).

These considerations can be extended to cases where additional information may be available, over and above the values of past X's, when each new X_i is to be forecast; then the place-selection rules entering the computable calibration requirement must also be allowed to take this information into account. Again it can be shown that all computably calibrated computable forecast systems must be in asymptotic agreement; but the implied "objective" probability forecasts will vary with the nature and extent of the available information. This conception of physical probability is thus a subtle and fundamentally relativistic one. In particular, there is no logical contradiction in believing in deep determinism (the case in which, for some suitably detailed information base, the asymptotically valid probability forecasts would all be 0 or 1), at the same time associating nonextreme probabilities with the outcomes at the less refined level of actually available information.

[13] The same conclusion must likewise hold when P and Q both pass the stronger martingale test of Section 4.1.

[14] Note, however, that no property anything like as strong as absolute continuity between P and Q is required here; in particular, nothing is assumed or asserted about whether P and Q agree, in any sense, for data-sequences other than the one actually observed.

5. CAUSALITY

Certain probabilistic scientific theories that we might consider could also contain other, nonprobabilistic, terms not relating directly to observables, and so not directly falsifiable. So long as we use such terms in a purely instrumental fashion, and regard as meaningful outputs of our theory only those assertions that do relate (according to our accepted rules of interpretation) to genuinely observable quantities, such additional terms can be accommodated in our general approach without handicap, and can sometimes simplify theoretical manipulations. However, care must be taken not to impart meaning to theoretical terms that do not correspond to observable quantities, and this is often harder than one might at first suppose. This problem arises in an especially acute form in the area of causal inference.

5.1 Potential Responses

Much current statistical work in causal inference lies within the framework of "potential response" modelling (Rubin, 1978; Dawid, 2000); a closely related approach is based on deterministic "structural relations" (Pearl, 2000; Dawid, 2002). A causal model constructed from such a standpoint might contain a term, Y_1 say, to represent "the duration of my current headache if I take aspirin" and another term, Y_0 say, for "its duration if I do not." It is clear what I have to do if I wish to observe either of these quantities: I either take, or refrain from taking, aspirin, and then I measure how long my headache lasts. In this sense, both Y_1 and Y_2 are (separately) empirically observable and meaningful. However, since, as a matter of logic, I can never simultaneously both take and not take aspirin, there is no conceivable measurement procedure that could evaluate, say, $Y_1 - Y_0$, which is thus a theoretical term with no physical counterpart. This lack of any link with the physical universe gives rise to grave philosophical and practical difficulties if, as is commonly done, one attempts to base causal inference on assertions about quantities such as $Y_1 - Y_0$. In particular, as pointed out by Dawid (2000), it is possible in these frameworks to set up distinct theoretical models that make identical assertions about all physically observable quantities—models that thus could never be distinguished on an empirical basis in any circumstances—but which imply different "causal conclusions" when these are phrased in terms of unobservable quantities such as $Y_1 - Y_0$. This behavior is in strong defiance of both Jeffreys's law and common

sense. These ambiguities cannot be removed except by imposing equally arbitrary and unfalsifiable additional constraints, such as "treatment-unit additivity." For these reasons I regard much of the current enterprise of causal inference as "fundamentally metaphysical." It cannot be disputed that much work of great value and importance has emerged, but that can largely be credited to the triumph of sound intuition over an inappropriate and cumbersome framework. And even a strong intuition may not be enough to locate unfailingly the delicate dividing line between those theoretical assertions that are empirically meaningful and those that (although equally well-formed mathematically) are scientific nonsense.

5.2 A Falsificationist Approach

Traditionally, statisticians shied away from any concern with causal analysis. Now that this enterprise is being taken more seriously, the prevalent view is that essentially new tools—such as joint modelling of potential responses—are required to take on this task. I do not subscribe to this view. I consider that the standard tools of probability and statistical theory are adequate for causal investigations—so long as the subtle relationships between model and reality, as already considered above, are kept clearly in mind. To paraphrase de Finetti: "Causality does not exist." That is, while we may find uses for causal terms within our theories in the intellectual universe, there are no direct external referents of such terms within the physical universe. Thus our understanding and interpretation of causal models of the world must again be indirect. However, no new ideas are needed: these tasks can be undertaken using the same falsificationist approach, based on Borel criteria, already developed for probabilistic modelling.

In my view, causal modelling and inference do not differ in any qualitative way from regular probabilistic modelling and inference. Rather, they differ quantitatively, because they have a more extended scope and ambition: namely, to identify and utilize relationships that are *stable* over a shifting range of environments ("*regimes*"). For example, we might entertain a collection of probability models for different economies, in different countries and different times, but relate these all together by assuming that they all share certain specified common probabilistic features—such as the distribution of the inflation rate, conditional on the central bank's base interest rate and the demand for and supply of manufactured goods. This is a causal theory, which might or might not be a good description.

Frequently, though by no means invariably, we are concerned with regimes that differ by virtue of possible *interventions* by an external agent, such as: give aspirin to, or withhold aspirin from, a patient; or recommend that the patient does or does not take aspirin; or simply observe whether or not the patient takes aspirin. For this reason we sometimes refer to this approach as "decision-theoretic," although this should not be understood as restricting its scope to interventional problems.

Often You will be able to conceive of many instances of each regime—instances that You would happily regard as exchangeable. For a generic instance of regime i there would be a multivariate collection $\mathbf{X}_i$ of observable quantities, with specific version $\mathbf{X}_{ij}$ for instance j of regime i. Under exchangeability across instances within each regime, an appropriate model would take all the $(\mathbf{X}_{ij})$ as independent, being identically distributed, with joint distribution P_i say, for all instances j of regime i—the (P_i) perhaps being subject to additional specification of their qualitative or quantitative structure and relationships. Such a model is falsifiable, using either the general approach of Section 3.3.4 or, in this case, the simpler frequentist approach of Section 3.3.3. We can interpret P_i in terms of limiting relative frequencies across many instances of regime i. Alternatively, if our model encompasses a potentially infinite collection of regimes, but only a finite number of instances of each, we could use a calibration-type Borel criterion to test its empirical validity.

A *causal model* imposes relationships between the various P_i describing the different regimes. For example, let $T = 1$ or 0 refer to the taking or not taking of aspirin, and Y to the duration of headache. In regime 1 (resp. 0) the patient is made to take (resp. not take) aspirin; in regime "$*$", the patient makes his own self-medication decision. We introduce conjectured joint probability distributions P_1, P_0 and P_* for (T, Y) in each of these regimes [for consistency of interpretation we require $P_1(T = 1) = P_0(T = 0) = 1$]. A possible causal model might then assert that the conditional distributions of Y given T are the same in all three regimes. Equivalently, using p_1 to represent a density under P_1, etc., we require $p_*(y|T = 1) = p_1(y)$, $p_*(y|T = 0) = p_0(y)$. That is to say, we are asserting that, so far as its predicted effect on Y is concerned, it does not matter how the treatment came to be administered. Such causal assertions are phrasable entirely within standard probabilistic language. They can be very naturally expressed and manipulated using

the notation and calculus of conditional independence, with or without associated graphical representations (Dawid, 2002); however, this is in no way fundamental.

The above model could be tested and possibly falsified in a straightforward fashion, by comparing the relevant relative frequencies across the three regimes. While there would often be severe practical impediments to gathering data under one or more of the regimes of interest, in principle at least it is clear what has to be done. This gives a clear empirical meaning to causal assumptions. When they can be regarded as valid, such assumptions can be used to support transfer of probabilistic information from one regime to another: an important purpose and use for causal analysis.

If we find that our causal model does not describe the physical world well, we might try and elaborate it until it does. One common approach is to try to identify some additional variable, U say, such that conditioning on U restores the desired invariance across regimes: $p_*(y|T=1,u) \equiv p_1(y|u)$, $p_*(y|T=0,u) \equiv p_0(y|u)$. Such a variable U is a "potential confounder." When U is unobserved or otherwise omitted, the invariance structure may well disappear, destroying or severely limiting the possibility of information transfer across regimes. An analysis of the problem of confounding from the decision-theoretic viewpoint can be found in Dawid (2002); Dawid (2003) studies what information can still be transferred in certain problems confounded by incomplete patient compliance with the doctor's treatment recommendations.

5.3 Causal Relativism

As is the case for our conception of Probability (see Section 4.2), an important feature of our approach to Causality is its relativism. We may choose to describe the world at various coarser or finer levels of detail. At any such level we might discover probabilistic invariances across different regimes, which we could term "causal." In some cases these might be reflections of causal relations (possibly deterministic, although this should not be assumed without good evidence) at a deeper level; but in general there need be no connections between the different levels of analysis. This means that a reductionist program, while often valuable and fruitful, cannot be universally successful: certain properties of coarse-grained systems may be genuinely emergent. Note that this understanding would simply not be available if, following most philosophical discussion of causation to date, we insisted on understanding all causal models in deterministic, rather than probabilistic, terms.

This minimalist decision-theoretic approach allows us to talk about causality in a simple common sense way, without needing to concern ourselves with issues of deep determinism or free will, or taking a fatalistic view of the world (Dawid, 2000). While philosophers may balk at its journeyman neglect of such issues, I feel that it liberates scientists to get on with doing useful things, without being sidetracked by metaphysical irrelevancies.

6. CONCLUSION

I have outlined a minimalist empiricist approach to Probability and Causality. In this view, probability and causality "do not exist." Rather, they are purely theoretical concepts, relating only indirectly to the physical universe. Probability assignments can be tested empirically, in the large rather than individually, by means of Borel falsifiability criteria such as calibration. Probability forecasts that are empirically valid attain a degree of objectivity: they are asymptotically unique (though typically unknowable). However, even this degree of objectivity remains relative, since the appropriate values depend on the nature and extent of the information available. Causal theories are nothing but ambitious probabilistic theories, positing certain common patterns of behavior across a range of different contexts, and raise no new issues of principle.

This program makes no metaphysical assumptions about the true nature or behavior of Probability or Causality, and is equally agnostic about such issues as deep determinism; but it does support a straightforward approach to building, testing, using and interpreting probabilistic theories of the world. When taken seriously, it necessitates some serious rethinking, or at least reinterpretation, of familiar patterns of argument: such as the assumption, underlying much of statistical inference, that there exists a "true (probabilistic) data-generating process" and that the task of inference is to learn something about it; or, in economic rational expectations theory (Muth, 1961) and game theory (Kalai and Lehrer, 1993), that an agent's or player's personal probability distribution over events should be the same as, or absolutely continuous with respect to, "Nature's true distribution." Although our approach is based on excluding from discussion any concept of a "true distribution," such hypotheses can be rephrased in our framework, as requiring agreement (as assessed by, e.g., calibration) between some probabilistic theory and actual *outcomes* in the world. Doing this then

enables us to consider more critically exactly what is being assumed or asserted by such a hypothesis, and thus how reasonable it might be.

REFERENCES

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598.

BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions* **53** 370–418 [Reprinted as Bayes (1958).]

BAYES, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika* **45** 293–315. [Reprint of Bayes (1763), with a biographical note by G. A. Barnard.]

BERNOULLI, J. (1713). *Ars Conjectandi*. Thurnisius, Basel.

BLACKWELL, D. and DUBINS, L. E. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33** 882–886.

BOREL, E. (1943). *Les Probabilités et la vie*. Presses Universitaires de France. [English translation published (1962) as *Probabilities and Life*. Dover, New York.]

CONSONNI, G. and DAWID, A. P. (1985). Invariant normal Bayesian linear models and experimental designs. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 629–643. North-Holland, Amsterdam.

COURNOT, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette, Paris.

DAWID, A. P. (1977). Invariant distributions and analysis of variance models. *Biometrika* **64** 291–297.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DAWID, A. P. (1982a). Intersubjective statistical models. In *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.) 217–232. North-Holland, Amsterdam.

DAWID, A. P. (1982b). The well-calibrated Bayesian (with discussion). *J. Amer. Statist. Assoc.* **77** 605–613. [Reprinted in Hamouda and Rowley (1997) 165–173.]

DAWID, A. P. (1984a). Causal inference from messy data. Discussion of "On the nature and discovery of structure," by J. W. Pratt and R. Schlaifer. *J. Amer. Statist. Assoc.* **79** 22–24. [Reprinted in Poirier (1994) **1** 368–370.]

DAWID, A. P. (1984b). Discussion of "Extreme point models in statistics," by S. L. Lauritzen. *Scand. J. Statist.* **11** 85.

DAWID, A. P. (1984c). Present position and potential developments: Some personal views. Statistical theory. The prequential approach (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

DAWID, A. P. (1985a). Calibration-based empirical probability (with discussion). *Ann. Statist.* **13** 1251–1285. [Reprinted in Hamouda and Rowley (1997) 174–208.]

DAWID, A. P. (1985b). The impossibility of inductive inference. Discussion of "Self-calibrating priors do not exist," by D. Oakes. *J. Amer. Statist. Assoc.* **80** 340–341.

DAWID, A. P. (1985c). Probability, symmetry and frequency. *British J. Philos. Sci.* **36** 107–128.

DAWID, A. P. (1986a). A Bayesian view of statistical modelling. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) 391–404. North-Holland, Amsterdam.

DAWID, A. P. (1986b). Probability forecasting. In *Encyclopedia of Statistical Sciences* **7** 210–218. Wiley, New York.

DAWID, A. P. (1988). Symmetry models and hypotheses for structured data layouts. *J. Roy. Statist. Soc. Ser. B* **50** 1–34.

DAWID, A. P. (1997). Prequential analysis. In *Encyclopedia of Statistical Sciences, Update Volume* **1** 464–470. Wiley, New York.

DAWID, A. P. (2000). Causal inference without counterfactuals (with discussion). *J. Amer. Statist. Assoc.* **95** 407–448.

DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Internat. Statist. Rev.* **70** 161–189. [Corrigenda (2002) **70** 437.]

DAWID, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with discussion). In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 45–81. Oxford Univ. Press.

DAWID, A. P. and VOVK, V. (1999). Prequential probability: Principles and properties. *Bernoulli* **5** 125–162.

DE FINETTI, B. (1937). Foresight: Its logical laws, its subjective sources. *Ann. Inst. H. Poincaré* **7** 1–68. [Translated by H. E. Kyburg in Kyburg and Smokler (1964) 93–158.]

DE FINETTI, B. (1974–1975). *Theory of Probability* **1**, **2**. Wiley, New York. [Italian original published (1970) by Einaudi, Torino.]

DIACONIS, P. and FREEDMAN, D. (1980). de Finetti's theorem for Markov chains. *Ann. Probab.* **8** 115–130.

GOODMAN, N. (1954). *Fact, Fiction, and Forecast*. Athlone, London.

HAMOUDA, O. F. and ROWLEY, J. C. R., eds. (1997). *Probability Concepts, Dialogue and Beliefs*. Elgar, Cheltenham, UK.

HOWSON, C. and URBACH, P. (1993). *Scientific Reasoning: The Bayesian Approach*, 2nd ed. Open Court, La Salle, IL.

KALAI, E. and LEHRER, E. (1993). Rational learning leads to Nash equilibrium. *Econometrica* **61** 1019–1045.

KYBURG, H. E. and SMOKLER, H. E., eds. (1964). *Studies in Subjective Probability*. Wiley, New York.

LAURITZEN, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics. Lecture Notes in Statist.* **49**. Springer, New York.

LAURITZEN, S. L. (2003). Rasch models with exchangeable rows and columns (with discussion). In *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 215–232. Oxford Univ. Press.

LINDLEY, D. V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge Univ. Press. (**1** Part I: Probability, **2** Part II: Inference.)

MILLER, R. I. and SANCHIRICO, C. W. (1999). The role of absolute continuity in "merging of opinions" and "rational learning." *Games Econom. Behav.* **29** 170–190.

MUTH, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica* **29** 315–335.

PEARL, J. (2000). *Causality*. Cambridge Univ. Press.

POIRIER, D. J., ed. (1994). *The Methodology of Econometrics* **1**, **2**. Elgar, Cheltenham, UK.

POPPER, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London. (German original published in 1934.)

POPPER, K. R. (1982). *The Open Universe*. Hutchinson, London.

POPPER, K. R. (1983). *Realism and the Aim of Science*. Hutchinson, London.

RAMSEY, F. P. (1926). Truth and probability. In *The Foundations of Mathematics and Other Logical Essays* (R. B. Braithwaite,

ed.) 58–100. Routledge and Kegan Paul, London. [Reprinted in Kyburg and Smokler (1964) 61–92.]

ROBINS, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS* (L. Sechrest, H. Freeman and A. Mulley, eds.) 113–159. NCSHR, U.S. Public Health Service.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

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

SCHERVISH, M. J. (1985). Discussion of "Calibration-based empirical probability," by A. P. Dawid. *Ann. Statist.* **13** 1274–1282.

SEILLIER-MOISEIWITSCH, F. and DAWID, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88** 355–359.

SHAFER, G. (1996). *The Art of Causal Conjecture*. MIT Press.

SHAFER, G. and VOVK, V. G. (2001). *Probability and Finance*: *It's Only a Game*. Wiley, New York.

VILLE, J. (1939). *Étude critique de la notion de collectif*. Gauthier-Villars, Paris.

VON MISES, R. (1939). *Probability, Statistics and Truth*. Hodge, London. [German original published (1928) by J. Springer.]