The Annals of Statistics
1985, Vol. 13, No. 4, 1251–1273

SPECIAL INVITED PAPER

CALIBRATION-BASED EMPIRICAL PROBABILITY

BY A. P. DAWID

University College London

Probability forecasts for a sequence of uncertain events may be compared with the outcomes of those events by means of a natural criterion of empirical validity, *calibration*. It is shown that any two sequences of forecasts which both meet this criterion must be in asymptotic agreement. These agreed values can then be considered as correct objective probability forecasts for the particular sequence of outcome results obtained. However, the objective forecasts vary with the extent of the information taken into account when they are formulated. We thus obtain a general theory of empirical probability, relative to an information base. This theory does not require that such probabilities be interpreted in terms of repeated trials of the same event. Some implications of this theory are discussed.

1. Introduction. Consider a probabilistic forecasting system that attaches numerical probabilities to each of a sequence of events. Each probability forecast is made only when the outcomes of previously forecast events have been determined. As examples, we might have a bookmaker who quotes odds on the favourite in a sequence of horse races, an economist who makes regular monthly probability forecasts of whether unemployment will rise or fall next month, a reliability engineer who gives probabilities of failure for some piece of equipment year by year, or a meteorologist who appears on television each evening and assigns a probability to the occurrence of rain in the area within the next 24 hours.

Dawid (1982a) considered the meaning that might be attached to a sequence of probability forecasts and introduced a criterion, calibration, which can be used to test the empirical validity of such a sequence in the light of the outcomes of the events being forecast. This paper investigates further those forecast sequences which are empirically valid by this criterion, and demonstrates that all of these must be in essential agreement, given sufficiently extensive experience. It thus follows that, for any empirical sequence of out-turns of the events, there must be an asymptotically *unique* acceptable "objective" sequence of values for the probabilities of the events (always conditional on previous experience). We thus obtain a powerful new generalization of traditional frequentist interpretations of probability, since we impose no requirement that the events under consideration be regarded as "unrelated trials under constant conditions" and allow the

Received September 1984; revised May 1985.

AMS 1980 *subject classifications*. Primary 60A05; secondary 03D10.

Key words and phrases. Calibration, empirical probability, forecasting system, prognostic system, computability, information base, inductive inference.

objective forecast probabilities to vary from day to day, as well as with the observed outcomes of past events.

The calibration criterion for an arbitrary sequence of probability forecasts in the light of a sequence of outcomes can, indeed, be considered as a natural extension of the attempt by von Mises, for the standard setting of repeated trials, to characterise an outcome sequence as "completely random." We briefly review this now classical theory in Section 2, which also motivates the introduction of the important idea of *computability* as fundamental to any rigorous development of such a characterization. Section 3 considers sequences that cannot simply be regarded as repeated trials, and formalises the general concept of a probability forecasting system. In Section 4 we discuss, in general terms, the kinds of property that are desirable in any criterion which purports to assess the empirical validity of such a probability forecasting system in the light of data. In particular, it is argued that asymptotic uniqueness should be such a desideratum. The next three sections introduce the calibration criterion, and show that, when suitably restricted by considerations of computability, it does possess the desired properties, including that of asymptotic uniqueness. In Section 8 we show that forecasts which are valid under this calibration criterion will minimise long-run average loss, as measured by a proper scoring rule.

All the considerations to this point relate to a forecaster whose information base at any time consists of the outcomes of previously forecast events. Sections 9 and 11 consider, respectively, the effects of expanding or restricting this information base. The calibration criterion again applies to such cases and again implies asymptotic uniqueness of valid forecasts. However, the "correct" valid probabilities will vary as we vary the information base. This is considered in Section 10. Section 12 extends the application of the previous criteria and results to the task of assigning a valid probability for an outcome event, as a function of specified covariate information. Finally the discussion in Section 13 points out some analogies with, and extensions of, the ideas developed in this paper.

2. Collectives. We begin by recapitulating the frequency theory of probability which grew out of the celebrated attempt of von Mises (1936) to define a random sequence. These ideas form a natural introduction to the more general theory to be introduced below. No attempt at completeness is made in this section. Those seeking a fuller account are referred to Martin-Løf (1969) or Knuth (1969, Section 3.5).

Let $\mathbf{a} = (a_1, a_2, \ldots)$ be an infinite sequence of 0s and 1s (sometimes we shall call these *failures* and *successes*, respectively), regarded as the outcomes of a sequence of trials. We seek to explicate the intuitive idea that this outcome sequence exhibits "randomness." A minimal requirement for this is generally accepted to be the existence of the *limiting relative frequency* $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} a_i = p$, say. Moreover, if we extract an infinite subsequence of a random sequence, this too should look random, with the same limiting relative frequency. To formalise this, let $s = (n_1, n_2, \ldots)$ be a subsequence of $(1, 2, \ldots)$, and define $\bar{a}_r(s) = r^{-1}\sum_{j=1}^{r} a_{n_j}$, the relative frequency of successes on the first r trials of the subsequence.

DEFINITION 2.1. We call **a** *invariant under s* if *either s* is finite *or* $\bar{a}_r(s) \to p$ $(r \to \infty)$. If $\mathscr{S}$ is a collection of subsequences, then **a** is *invariant under* $\mathscr{S}$ if it is invariant under every $s \in \mathscr{S}$.

Intuitively, a random sequence should be invariant under any subsequence s. However, we cannot sensibly require invariance under the (uncountable) collection of *all* subsequences. For instance, one such subsequence will consist of just those values n for which $a_n = 0$, and so yield $\bar{a}_r(s) \equiv 0$. A way must be found to exclude such "peculiar" subsequences.

One possibility is to settle in advance, arbitrarily, on some countable collection $\mathscr{S}$ of subsequences and to require invariance under this $\mathscr{S}$. However, this restriction errs too far to the other extreme. It seems intuitively reasonable that we should require invariance under certain *data-dependent* subsequences (for example, the subsequence of all trials following a failure). If invariance did not hold under this subsequence, that would appear to indicate a dependence between the outcomes of successive trials. But any member of the uncountable collection of all subsequences might be the result of such a process for some outcome sequence **a**. Thus a different approach is suggested.

DEFINITION 2.2. A *selection rule* is a function from the set of all finite (possibly empty) strings of 0s and 1s into the set $\{0, 1\}$.

If **a** is a sequence of outcomes, $\mathbf{a}^{(n)} = (a_1, a_2, \ldots, a_n)$ will denote its initial string of length n. Then a selection rule δ will be said to *select the subsequence* $s = (n_1, n_2, \ldots)$ *under* **a** if $n \in s$ just when $\delta(\mathbf{a}^{(n-1)}) = 1$. In that case we may write $s = \delta(\mathbf{a})$.

Informally, the trials selected by a selection rule are determined only by outcomes of previous trials. If, as randomness suggests, there are no "carry-over" effects in the sequence, then the selected outcomes should still look random. In other words we should require **a** to be invariant under the sequence $\delta(\mathbf{a})$, for any selection rule δ. This is von Mises' randomness criterion and he termed a sequence **a** satisfying it a *collective*.

Unfortunately, as noted by Wald (1937), this definition will not do, for exactly the same reason that we cannot require invariance under every fixed subsequence. Indeed, given any "peculiar" subsequence s, such as that of all failures in **a**, we can trivially produce a selection rule δ which selects s under any outcome sequence whatsoever. We have again tipped the balance too far.

As a reasonably satisfactory compromise position, we might finally settle on some *countable* collection $\mathscr{C}$ of *selection rules*. Then **a** is called a *collective with respect to* $\mathscr{C}$, or *$\mathscr{C}$-collective*, if it is invariant under $\mathscr{C}(\mathbf{a}) = \{\delta(\mathbf{a}): \delta \in \mathscr{C}\}$. This is a countable collection of subsequences for any fixed **a**, but the collection itself varies with **a**. Wald showed the consistency of this definition. Indeed, if we regard **a** as the outcomes of Bernoulli trials with probability p, then, for any $\mathscr{C}$, **a** will satisfy this definition with probability 1. Had this result been false, we might have had serious doubts that we had truly captured the idea of randomness. (It is still possible to argue, as Ville (1939) has done, that the criterion of being a

collective is too weak, and that further conditions should be imposed. See Section 13.2.) There remains the problem of how to choose $\mathscr{C}$.

DEFINITION 2.3. A selection rule δ is *computable* if there exists a Turing machine which, when fed with an input tape containing any finite (possibly empty) string of 0s and 1s, stops after a finite number of operations having output the value of δ for that string.

Since any abstract Turing machine can be finitely described, there is only a countable number of such machines. Thus, we can, and shall, following Church (1940), take for $\mathscr{C}$ the collection $\mathscr{C}^*$ of all computable selection rules. This is the broadest class with which we can meaningfully work, since it may plausibly be supposed that any selection rule that can be humanly constructed will be computable. (It is perhaps worth noting that the original definition of a Turing machine (Turing, 1936) was the result of a deliberate attempt at modelling human information processing.)

We have thus arrived, finally, at the concept of the *$\mathscr{C}^*$-collective*, a criterion which applies to an infinite sequence $\mathbf{a}$, and which, when it is satisfied, can be interpreted as asserting that $\mathbf{a}$ looks "random with probability parameter p." Henceforth, we shall simply call such a sequence $\mathbf{a}$ a *collective*.

For further developments, it will be helpful to reinterpret this criterion slightly. Given an empirical outcome sequence $\mathbf{a}$, we can consider, as a possible probabilistic explanation of $\mathbf{a}$, the model of *Bernoulli trials* with probability parameter p. We can then regard this as acceptable if, and only if, $\mathbf{a}$ forms a collective (with this probability parameter). Thus the criterion is now regarded as applying to the hypothetical Bernoulli probability model, rather than to the data. If and only if it is satisfied, the Bernoulli model can be accepted as an "explanation" of the data.

3. Forecasting systems. We now wish to extend the idea that we might judge the suitability of a hypothetical probabilistic model as an explanation of given data, beyond the special case of the Bernoulli model. We shall recast this goal as the evaluation of *forecasting systems*.

Let $\mathbf{A} = (A_1, A_2, \dots)$ denote an infinite sequence of uncertain events, identified with their indicators so that $A_n = 1$ if the nth event occurs, $A_n = 0$, otherwise. We denote $(A_1, A_2, \dots, A_n)$ by $\mathbf{A}^{(n)}$, and the σ-field generated by $\mathbf{A}^{(n)}$ (resp. $\mathbf{A}$) by $\mathscr{A}_n$ (resp. $\mathscr{A}_\infty$).

We imagine the outcomes $\mathbf{a} = (a_1, a_2, \dots)$ of $\mathbf{A}$ to be observed sequentially. After observing the outcomes, $\mathbf{A}^{(n)} = \mathbf{a}^{(n)}$, of the first n trials, it is required that a probability p_{n+1} be assigned to the occurrence of the next event A_{n+1}. Any method of constructing such *probability forecasts*, for every n and $\mathbf{a}^{(n)}$, will be called a *forecasting system*. Such a forecasting system is thus a function (F say) from the set of finite (possibly empty) strings of 0s and 1s, into the interval $[0, 1]$ (compare Definition 2.2). Its value $F(\mathbf{a}^{(n)})$ for argument $\mathbf{a}^{(n)}$ is taken as p_{n+1}. If $\mathbf{p}$ is the corresponding infinite forecast sequence $(p_1, p_2, \dots)$, we shall also write $\mathbf{p} = F(\mathbf{a})$.

If Π is any probability model over $\mathscr{A}_\infty$, then Π determines a forecasting system F, by $F(\mathbf{a}^{(n)}) = \Pi(A_{n+1}|\mathbf{A}^{(n)} = \mathbf{a}^{(n)})$. (We assume throughout, for simplicity, that no finite string of outcomes is assigned probability 0 by Π.) When Π is the Bernoulli trials model with parameter p, the corresponding F yields the constant forecasts $F(\mathbf{a}^{(n)}) \equiv p$, for any n and $\mathbf{A}^{(n)}$.

Conversely, given any forecasting system F, there exists a unique distribution Π giving rise to it in this way.

Forecasting systems may, however, be derived in many other ways (Dawid, 1984). For example, we might specify a parametric family $\mathscr{P} = \{P_\theta\}$ of joint distributions for $(A_1, A_2, \ldots)$, and take $P_{n+1} = P_{\hat{\theta}_n}(A_{n+1}|A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$, where $\hat{\theta}_n$ is, say, the maximum likelihood estimator of θ based on the observed outcomes $\mathbf{a}^{(n)}$. We could also consider "supersystems" of the following kind. A specific forecasting system is, initially, laid down and used. At periodic intervals, some sort of comparison is made between past forecasts and the out-turn of events, perhaps by means of a suitable significance test. One approach to such a comparison might be based, for example, on the calibration criterion of Section 5, suitably interpreted for finite outcome sequences. If the outcome of such a comparison is unsatisfactory, the initial forecasting system might be modified or replaced by a new one, perhaps attempting to take account of previously unsuspected patterns discovered in the data. This whole process can be repeated at regular or irregular intervals. Such a supersystem embodies the spirit of the recommendations of Box (1980), in which regular periods of "estimation" (use of a particular system) are interspersed with bouts of "criticism" (leading, possibly, to the overthrow of the old system and the rise of a new one). It is also close in spirit to the alternation of "normal science" and "scientific revolutions" conceived by Kuhn (1962). One can even turn this approach in on itself, and consider superdupersystems, whose basic building blocks are supersystems, which are replaced when they no longer work. And so on, through an endless ordinal sequence. But, when all this has been done, one is still left with a single final forecasting system F, which can be evaluated like any other.

Lastly, we might also admit still more informal forecasting systems, where a meteorologist (for example) gives a "seat-of-the-pants" subjective probability forecast in the light of his or her background information (Dawid, 1985). (For cases in which additional information over and above past outcomes is being used, see Section 9.)

4. Metacriteria. Suppose that Nature determines a specific realised sequence $\mathbf{a}$ of outcomes of $\mathbf{A}$. Then a forecasting system F will produce the string $\mathbf{p} = F(\mathbf{a})$ of probability forecasts. If possible, we wish to examine the success of F in explaining the specific outcomes in $\mathbf{a}$.

The philosophy underlying the approach to be taken is that stochasticity should be regarded as an attribute, not of any external process which generates the outcomes $\mathbf{a}$, but of the hypothetical probability models or forecasting systems proposed as possible explanations of $\mathbf{a}$. If such a model provides (in a sense yet to be made precise) a "successful," or "(empirically) valid" explanation, we may conclude that $\mathbf{a}$ "looks like" it was generated from that model, but we should *not*

conclude that we have identified "the true model" governing Nature's production of **a**. Indeed, there is no evident reason why two different models should not both provide successful explanations of the same data **a**. This approach, which proceeds by pitting hypothetical explanations against empirically observed data, is in close accord with the general scientific methodology of Karl Popper. By taking an entirely instrumental approach to probability modelling, however, it dispenses with any need for a "realist" interpretation of probability.

While it is not immediately clear what assessment criterion should be used to judge the success of F in explaining **a**, the following *metacriteria* for choosing such a criterion may be considered more or less compelling.

M1. The criterion should be applicable to any forecasting system F and data sequence **a**. When F corresponds to the Bernoulli model, it should reduce to an accepted criterion for the "randomness' of the sequence **a**.

M2. The criterion should depend only on **a**, the realised outcomes, and **p**, the forecasts actually made by F.

M3. If Π is a distribution over $\mathscr{A}_\infty$, giving rise to a forecasting system F, then the set of outcome sequences for which F is a valid explanation should have probability 1 under Π.

M4. If F^1 and F^2 are both valid explanations of **a**, with corresponding forecast sequences $\mathbf{p}^1$ and $\mathbf{p}^2$, then $p_n^1 - p_n^2$ should tend to zero as $n \to \infty$.

Notes on the metacriteria. M1 is an expression of the broad basis of our approach, that empirically valid probabilities should be meaningful even in the absence of any setup of "repeated trials under constant conditions" (although that setup must be seen as an important special case).

M2 is intuitively appealing, since we wish to assess how well F has performed in this world, not in hypothetical worlds which have not materialised. It is also necessary if we are to be able to assess more informal forecasting systems, such as that of the meteorologist who only quotes his probability of rain tomorrow in the light of actual, not hypothetical, past data, or an incompletely formalised supersystem. Also, by shunning any recourse to hypothetical repetitions (of the whole event sequence), M2 draws still further away from the unnecessary and restrictive idea that empirical probability can only be understood in terms of such repetitions.

M3 requires that a probability model should be a valid explanation of almost all the sequences that arise from that model. (This too has been generally accepted as a metacriterion for criteria of "randomness" in relation to Bernoulli trials.) If M3 failed, we might reasonably regard our criterion as too strong. If two different criteria both satisfy M3, they can both be expected to return the same verdict as to whether F is a valid explanation of **a**, for "most" sequences **a** (but see the caution under Theorem 4.1).

Note, however, that M3 limits the extent to which we can discriminate between different probability models. Thus let Π^1 and Π^2 be distributions over $\mathscr{A}_\infty$ with $\Pi^1 \ll \Pi^2$ (i.e., Π^1 is *absolutely continuous* with respect to Π^2, so that if $S \in \mathscr{A}_\infty$ and $\Pi^2(S) = 0$, then $\Pi^1(S) = 0$). For $i = 1, 2$, let S^i be the event that, for some given criterion satisfying M3, the corresponding forecasting system

F^i is a valid explanation of the realised outcome sequence. Then, by M3, $\Pi^1(S^1) = 1$ and $\Pi^2(S^2) = 1$. In particular, both F^1 and F^2 will be regarded as valid whenever $S^0 = S^1 \cap S^2$ obtains, an event which has Π^1-probability 1. Thus we cannot expect to distinguish between F^1 and F^2, by any such criterion, for a large class of outcome sequences.

With this ambiguity in mind, it might be thought, initially, that M4 is too strong and inconsistent with the other metacriteria. Indeed, the present investigation arose from the author's original belief that this was so and that many essentially different forecast sequences might reasonably be considered as equally valid explanations of a given outcome sequence **a**. For example, if two weather forecasters persist in producing quite different day-by-day probability forecasts for rain, can we not allow that, though different, both sequences might turn out to be valid probabilistic explanations of the weather, particularly in the light of the arbitrariness shown to be inherent in M3 above? However, the following result gives some indication that M4 may not, after all, be too strong.

THEOREM 4.1 (Blackwell and Dubins, 1962). *Let* Π^1 *and* Π^2 *be distributions over* $\mathscr{A}_\infty$ *with* $\Pi^1 \ll \Pi^2$. *Then with* Π^1*-probability* 1, $\sup\{|\Pi^1(S|\mathscr{A}_n) - \Pi^2(A|\mathscr{A}_n)|\colon S \in \mathscr{A}_\infty\} \to 0$ $(n \to \infty)$.

COROLLARY. *If* $P^i_{n+1} = \Pi^i(A_{n+1}|\mathscr{A}_n)$, *then, with* Π^1*-probability* 1, $P^1_n - P^2_n \to 0$ $(n \to \infty)$.

In other words, for "most" outcome sequences for which both Π^1 and Π^2 provide valid explanations, their forecast sequences will be asymptotically indistinguishable, as required by M4. However, this result can be no more than suggestive: it is a long way from "most" to "all."

As we shall see, it is indeed possible to produce criteria for which all of the metacriteria M1–M4 hold, and this, once being shown possible, may then reasonably be regarded as obligatory.

Notice that M4 does *not* constrain F^1 and F^2 to be asymptotically indistinguishable as forecasting systems. In the spirit of M2, it is only required that $p^1_n - p^2_n \to 0$ for the actual data sequence **a** obtained, not necessarily for other hypothetical sequences. It will thus not be possible (nor, I believe, desirable) to distinguish between different forecasting systems or models which just happen to yield identical forecasts for the outcomes which Nature produces. In this sense, M4 does not justify us in calling an empirically valid model "the true model", but it *does* justify us in considering a valid sequence of *forecasts* as "true," or "objective," at least asymptotically. Thus a criterion satisfying M1–M4 can be said to yield an empirical concept of probability, of great generality, which comes as close as may be reasonably expected to justifying unique "correct" probabilities for individual events.

A consequence of M3 is that no definitive rejection of F can be made from any finite string of outcomes (as long, at any rate, as we continue with our simplifying assumption that such a string is assigned positive probability). Similarly (even without such an assumption), M4 implies that no definitive acceptance is possible

with only finitely many data. In other words, under M3 and M4, probabilities can only be validated in infinite sequences. While this creates a serious (some might say insurmountable) obstacle to practical implementation of any validity criterion, it is surely the very essence of any reasonable empirical interpretation of probability in terms of a sequence of data outcomes. Certainly, it holds for any traditional approach based on limiting relative frequencies.

In this connexion, note that it is impossible to strengthen M4 to the point that it insists on individually unique, rather than asymptotically unique, valid probability forecasts. Indeed, it is an essential limitation of *any* frequency theory of empirical probability based on limiting properties of infinite sequences that we can alter the probabilities attached to an arbitrarily large, but finite, set of events, without disturbing any of those limiting properties. Consequently, no such theory can ever justify assigning particular probabilities to particular events. The most that can be expected is that some asymptotic assignment may be justified and it is in this spirit that we propose M4.

5. Calibration. We shall now exhibit a criterion which satisfies M1–M4, and thus justifies asymptotically unique "objective" probability forecasts for events. It is a natural extension of the ideas in Section 2, which related to Bernoulli trials only.

For a subsequence $s = (n_1, n_2, \ldots)$, denote by $\bar{p}_r(s)$ the average probability forecast, $r^{-1}\sum_{j=1}^{r} p_{n_j}$, for the first r events in s.

DEFINITION 5.1. We say that **p** is *calibrated for* **a** *with respect to* s if *either* s is finite *or* $\bar{a}_r(s) - \bar{p}_r(s) \to 0$ $(r \to \infty)$.

If δ is a selection rule, we say that **p** *is calibrated for* **a** *with respect to* δ if calibration holds with respect to the subsequence $s = \delta(\mathbf{a})$ selected by δ under **a**.

If $\mathscr{C}$ is a collection of selection rules, we say that **p** is *completely calibrated for* **a** *with respect to* $\mathscr{C}$ if calibration holds with respect to every $\delta \in \mathscr{C}$.

Finally, if **p** is completely calibrated for **a** with respect to the class $\mathscr{C}^*$ of all computable selection rules, we shall call **p** *computably calibrated* for **a**.

The intuitive concept of calibration is that, for all suitably specified subsequences, the probability forecasts should be right "on average" in comparison with relative frequencies, at any rate asymptotically. Further, if the quoted probability forecasts truly do take full and correct account of all previous outcomes, they should remain appropriate even for events picked out on the basis of those outcomes, which is why we must require calibration with respect to selection rules. Indeed, such a selection rule may fruitfully be regarded as a strategy selected by an adversary who is trying to discredit **p** as a valid sequence of forecasts, by picking some subsequence for which he believes the forecasts are inappropriate (for example, too optimistic). We must allow such an adversary access to the same past data as the forecaster, to give him scope to show that the forecaster has not taken correct account of those data. A forecast sequence cannot be regarded as valid if such an adversary can prove his case by showing that **p** is not calibrated for **a** with respect to his selected subsequence.

One possible selection rule might pick out all those events A_n for which the forecasting system F under test assigned probability p_n satisfying, for example, $|p_n - 0.4| < 0.05$. This does indeed depend only on the outcomes $\mathbf{a}^{(n-1)}$ of previous events, since this property is true of $p_n = F(\mathbf{a}^{(n-1)})$. If the selected subsequence is infinite, $\mathbf{p}$ will be calibrated for $\mathbf{a}$ with respect to this selection rule only if the limiting relative frequency of success, among the events whose probability forecast is 0.4 (to one decimal place), is itself 0.4 (to the same accuracy). An identical conclusion holds for any other "target probability" and accuracy. Traditionally, the calibration criterion has been applied only for such selection rules, which group events according to their assigned probabilities (see e.g., Lichtenstein et al., 1982). However, Definition 5.1 imposes no such restriction, and complete calibration is consequently a much more stringent condition on $\mathbf{p}$ than traditional calibration. For example, one could use a selection rule picking out only a subsubsequence of that for which $|p_n - 0.4| < 0.05$, by imposing additional inclusion criteria depending on previous outcomes. In particular, calibration under an appropriate collection of such rules would ensure that the outcomes of all the events assigned probability (about) 0.4, say, should be a *collective* with that probability parameter, as suggested by Curtiss (1968). (Even this condition, however, is less stringent than complete calibration.)

The sequence of definitions in Definition 5.1 parallels exactly the developments described in Section 2, and its culmination in computable calibration follows from the same logical considerations that led to the representation of randomness by the $(\mathscr{C}^*)$-collective. Henceforth, we shall take computable calibration as our criterion for the validity of $\mathbf{p}$ as an explanation of the data sequence $\mathbf{a}$. We shall justify this choice by showing that it satisfies the metacriteria M1–M4.

It is easy to see that M2 holds for this criterion. For M1, we note that, since the Bernoulli forecasting system has $p_n \equiv p$, then $\bar{p}_r(s) \equiv p$, so that we recover in that case the criterion that $\mathbf{a}$ should form a collective, our accepted criterion for randomness.

That M3 holds is less obvious, but does indeed follow from the theorem of Dawid (1982a), which asserts that, with Π-probability 1, the outcome sequence $\mathbf{a}$ will be such that the corresponding forecasts $\mathbf{p}$ produced by F will be calibrated for $\mathbf{a}$ with respect to any prespecified selection rule. Thus our criterion is not too strong.

It remains to investigate M4. To see that some caution is required here, consider, for example, an outcome sequence $\mathbf{a}$ forming a collective, with probability parameter p, and the following two sequences of forecasts, $\mathbf{p}^1$ and $\mathbf{p}^2$. For $\mathbf{p}^1$ we have $p_n^1 \equiv p$, according to the usual Bernoulli model. For $\mathbf{p}^2$ we have $p_n^2 = a_n$, corresponding to perfect forecasting with certainty. It is clear that both these forecast sequences are computably calibrated for $\mathbf{a}$, but $p_n^1 - p_n^2 \not\to 0$ in contradiction to M4.

Intuitively, it seems clear that the forecasts $\mathbf{p}^2$ should be ruled out, not because they fail to satisfy our criterion, but because they are "too good." It is simply impossible to forecast a collective perfectly, at any rate in the absence of clairvoyance or additional external information. We must therefore impose some such constraint on the forecasts we may consider. This idea is taken up in the next section; we then return to investigate M4 in Section 7.

6. Computable forecast sequences. Our restriction on the allowable forecast sequences will parallel closely that already imposed on selection rules, that they be, in a suitable sense, computable. For our present purposes, a fairly weak definition will suffice.

DEFINITION 6.1. A forecasting system F is called *simply computable* if its value for any string has the form $a2^{-b}$, for integral a and b, and there exists a Turing machine, which, when fed with an input tape containing any finite (possibly empty) string of 0s and 1s, stops after a finite number of operations, having output the (finite) binary expansion of F for that string.

Given a specified outcome sequence $\mathbf{a}$, a forecast sequence $\mathbf{p}$ is called *simply computable* for $\mathbf{a}$ if there exists a simply computable forecasting system F such that $\mathbf{p} = F(\mathbf{a})$, viz. $p_{n+1} = F(\mathbf{a}^{(n)})$, all n. We call $\mathbf{p}$ *computable* for $\mathbf{a}$ if there exists a simply computable *approximating* forecast sequence $\mathbf{q}$ for $\mathbf{a}$, viz. one such that $p_n - q_n \to 0$ $(n \to \infty)$.

This author would again hold that, if we restrict attention to computable forecast sequences, we shall not have excluded any that are humanly attainable (still excluding clairvoyance or the possession of additional external information). Without entering into any further philosophical debate on this point, we now impose this restriction. With it, the possibility of perfect forecasting of a collective, for example, is ruled out. If it were possible, there would exist a computable selection rule which picked out all the events, beyond some point in the sequence, which result in a success, in contradiction to the requirement of invariance under this subsequence.

7. Asymptotic uniqueness. We now present the principal result of this investigation that, when only computable forecast sequences are admitted, M4 holds for the criterion of computable calibration. (Note that, if $\mathbf{p}^1$ is empirically valid under these conditions and $p_n^1 - p_n^2 \to 0$, then $\mathbf{p}^2$ is empirically valid. Hence M4 is the strongest such requirement that can be imposed for this criterion.)

LEMMA 7.1. *Let $\mathbf{p}^1, \mathbf{p}^2$ be forecast sequences, each calibrated for $\mathbf{a}$ with respect to a subsequence s. Then, if s is infinite, $\bar{p}_r^1(s) - \bar{p}_r^2(s) \to 0$ $(r \to \infty)$.*

PROOF. Immediate from Definition 5.1. □

THEOREM 7.1. *Let $\mathbf{p}^1, \mathbf{p}^2$ be computable forecast sequences for $\mathbf{a}$, each computably calibrated for $\mathbf{a}$. Then $p_n^1 - p_n^2 \to 0$ $(n \to \infty)$.*

PROOF. Clearly it is enough to prove the result when each $\mathbf{p}^i$ is simply computable, so that $\mathbf{p}^i = F^i(\mathbf{a})$ for some simply computable forecasting system F^i $(i = 1, 2)$. For this case, given any integer K, we can define a selection rule δ_K by: $\delta_K(\boldsymbol{\alpha}^{(n)}) = 1$ if $F^1(\boldsymbol{\alpha}^{(n)}) - F^2(\boldsymbol{\alpha}^{(n)}) > K^{-1}$; $\delta_K(\boldsymbol{\alpha}^{(n)}) = 0$, otherwise. Furthermore, δ_K is computable, since a Turing machine can be constructed which combines those which compute F^1 and F^2 with a further step which checks

whether or not the condition $F^1(\boldsymbol{\alpha}^{(n)}) - F^2(\boldsymbol{\alpha}^{(n)}) > K^{-1}$ is satisfied. It follows that both $\mathbf{p}^1$ and $\mathbf{p}^2$ must be calibrated for $\mathbf{a}$ with respect to the subsequence $s_K = \delta_K(\mathbf{a})$, for which $n \in s_K$ if and only if $p_n^1 - p_n^2 > K^{-1}$. Then, for all r, $\bar{p}_r^1(s_K) - \bar{p}_r^2(s_K) > K^{-1}$. Thus s_K must be finite, for, if not, Lemma 7.1 would yield a contradiction. The same argument applies on interchanging F^1 and F^2. We have thus shown that, for any integer K there exists N such that $|p_n^1 - p_n^2| \leq K^{-1}$ for all $n > N$, i.e., $p_n^1 - p_n^2 \to 0$. □

Note 7.1: Universal algorithm. Theorem 7.1 essentially states that, under the calibration criterion, if valid probability forecasts exist at all, then they are, asymptotically, uniquely determined. The demonstration is, however, disappointingly nonconstructive. It is tempting to think that there might exist a "universal algorithm" that could sequentially process the data and construct the valid forecast sequence when it exists. Unfortunately, this assumption may be shown to generate a contradiction (Oakes, 1985). There is no computable way of discovering what the "correct" forecast sequence is!

It is instructive to examine the following hopeful attempt to construct a universal algorithm. Let $F_1, F_2, \ldots$ be a complete listing of the countable collection of simply computable forecasting systems, with corresponding distributions $\Pi_1, \Pi_2, \ldots$. Define $\Pi_0 = \sum_{i=1}^{\infty} \Pi_i / 2^i$. Then $\Pi_i \ll \Pi_0$. There exist sets of sequences, $S_1, S_2, \ldots$, such that $\Pi_i(S_i) = 1$, F_i is computably calibrated for any $\mathbf{a} \in S_i$, and (by the corollary to Theorem 4.1) so too are the forecasts made by Π_0. Thus Π_0 will be computably calibrated for any $\mathbf{a} \in S = \bigcup_{i=1}^{\infty} S_i$, a set which has probability one under *any* computable distribution Π_i. This would seem to come as near as one might require to providing a universal algorithm.

However, Π_0 above does *not* correspond to a computable forecasting system. The above programme cannot be carried out in an effective manner, since it is impossible to order the (F_i) effectively. If we could do so, the rule $F^*(\mathbf{a}^{(n)}) = 1 - F_n(\mathbf{a}^{(n)})$ if this is not $\frac{1}{2}$, or $\frac{3}{4}$ if it is, would determine a simply computable forecasting system that differs from all the (F_i). [This diagonal argument goes back, in essence, to the original paper of Turing (1936).]

Note 7.2: Stable estimation. A well-known result in Bayesian inference (see, for example, Edwards et al., 1963) asserts, informally, that if two forecasters agree on the form of a statistical model for data $\mathbf{X} = (X_1, X_2, \ldots)$ given a parameter θ, but have different prior distributions for θ, then, as $n \to \infty$, their posterior distributions for θ given $\mathbf{X}^{(n)} = (X_1, X_2, \ldots, X_n)$ will tend to agreement. So too, it follows, will their predictive distributions for X_{n+1} given $\mathbf{X}^{(n)}$. For some models at least, this asymptotic agreement will hold for *any* sequence of outcomes.

The Blackwell–Dubins result (Theorem 4.1) can be regarded as an extension of this argument when the two forecasters do not necessarily share a common model, but do at least have mutually absolutely continuous distributions, and so agree with each other on what events are to be regarded as certain or impossible.

Then they will be in asymptotic agreement for a set of outcome sequences to which each attaches probability 1.

Neither of these arguments implies that the resulting agreed predictions will be empirically meaningful in any way. However, Theorem 7.1 demonstrates that, if two forecasters do both succeed in making empirically valid forecasts, then they must be in asymptotic agreement. Neither need base his forecasts on any specific modelling assumptions, nor, if they do, need they agree over these, nor need their overall distributions be mutually absolutely continuous.

Note 7.3: Calibrability. Theorem 7.1 does not guarantee that, given an outcome sequence **a**, there will necessarily exist any empirically valid forecast sequence **p**. If this holds, we may call **a** *calibrable*. From M3, we know that the set of noncalibrable sequences has probability 0 under any probability distribution that corresponds to a computable forecasting system, and so is evidently very sparse in an intuitive sense. It is not empty, however. Schervish (1985) has shown that there exist uncountably many noncalibrable sequences. A similar conclusion has been reached by, among others, Sudbury (1973).

If **a** is a calibrabie sequence, a valid forecast sequence **p** will generally be nondegenerate, in the sense that $p_n(1 - p_n)$ does not tend to zero. This will certainly be the case for almost all realisations from a distribution Π having this property for $p_n = \Pi(A_n|A_1, A_2, \ldots, A_{n-1})$. In particular, such a sequence **a** cannot itself be computable, for, if it were, we could take the perfect forecasts $\mathbf{p} = \mathbf{a}$ as empirically valid in contradiction to Theorem 7.1. Thus calibrable sequences are themselves highly irregular, but the very fact of calibrability implies some deeper underlying regularity. For a noncalibrable sequence, even this "order in chaos" is absent.

8. Scoring rules. A popular way of evaluating individual probability forecasts is by means of a *scoring rule* (Savage, 1971): a function $S = S(a, p)$ of the outcome a of the event A being forecast and the quoted forecast probability p. We regard this as a penalty to be paid. If a forecaster's "true" probability of A is p_0, then his expected score, if he quotes p, is $S(p_0, p) = p_0 S(1, p) + (1 - p_0)S(0, p)$. If this is (uniquely) minimised in p for $p = p_0$, the scoring rule is called (*strictly*) *proper*. Thus a proper scoring rule motivates the forecaster truthfully to quote his true probability. An arbitrary decision problem, with loss function $L(a, d)$ depending on the decision d and outcome a of A, determines a proper scoring rule S, such that $S(a, p)$ is the loss suffered by taking the decision d_p optimal under the quoted probability p for A, when $A = a$. If the forecaster now faces a sequence **A** of such events, with the same decision problem at each stage, it seems reasonable to measure the badness of his quoted probability forecasts **p**, in the light of outcomes **a**, by the average loss which they imply (Dawid, 1985a). This motivates the following.

DEFINITION 8.1. Let S be a proper scoring rule. A sequence **p** of probability forecasts is said to be *S-superior* to another such sequence **q**, with respect to the

outcome sequence **a**, if

$$\liminf_{n\to\infty} n^{-1} \sum_{i=1}^{n} \{S(a_i, q_i) - S(a_i, p_i)\} \geq 0.$$

We might consider, as another possible validity criterion for **p**, the requirement that **p** be S-superior to any other contending forecast sequence. The following result shows that, under a continuity condition, this is implied by our criterion of computable calibration.

THEOREM 8.1. *Let $S(a, p)$ be a proper scoring rule, continuous in $p \in [0,1]$ $(a = 0,1)$. If forecast sequences* **p** *and* **q** *are computable for the outcome sequence* **a**, *and, moreover,* **p** *is computably calibrated for* **a**, *then* **p** *is S-superior to* **q**.

PROOF. Since continuity on a compact interval implies uniform continuity, it is easily seen that the result will hold if it holds for simply computable forecast sequences approximating **p** and **q**. Thus we now suppose **p** and **q** to be simply computable.

Let $y_i = S(a_i, q_i) - S(a_i, p_i)$, $\bar{y}_n = n^{-1}\Sigma_{i=1}^{n} y_i$. We must show that $\liminf_{n\to\infty} \bar{y}_n \geq 0$.

Given $\varepsilon > 0$, we can find rational $\delta > 0$ such that, for $p, p' \in [0,1]$, $|p - p'| \leq \delta \Rightarrow |S(r, p) - S(r, p')| \leq \varepsilon$ for $r = 0$ or 1 and, thus, for any $r \in [0,1]$. We may then partition the interval $[0,1]$ into a finite number N of intervals $(I_1, I_2, \ldots, I_N)$, each of length $\leq \delta$, and such that each I_k has rational upper and lower endpoints u_k and l_k, with $u_k = l_{k+1}$ $(k = 1, \ldots, N-1)$. We may correspondingly partition the unit square into N^2 boxes $(B_{jk}: 1 \leq j, k \leq N)$, where $B_{jk} = I_j \times I_k$; and the sequence $(1, 2, \ldots)$ into N^2 corresponding subsequences (s_{jk}), where $i \in s_{jk}$ if and only if $(p_i, q_i) \in B_{jk}$. Note that each such subsequence is computable. We denote by s_{jk}^n the intersection of s_{jk} with $\{1, 2, \ldots, n\}$, and by n_{jk} the size of this set.

Define $z_i = S(a_i, l_k) - S(a_i, u_j)$ if $i \in s_{jk}$ with $j < k - 1$, $z_i = S(a_i, u_k) - S(a_i, l_j)$ if $i \in s_{jk}$ with $j > k + 1$, and $z_i = 0$ if $i \in s_{jk}$ with $|j - k| \leq 1$. Then, for all i, $|y_i - z_i| \leq 2\varepsilon$, so that $|\bar{y}_n - \bar{z}_n| \leq 2\varepsilon$.

Let $\bar{z}_{jk}^n$ (resp. $\bar{a}_{jk}^n$) denote the average of the z_i (resp. a_i) for $i \in s_{jk}^n$. Then $\bar{z}_n = \Sigma_{j=1}^{N}\Sigma_{k=1}^{N}(n_{jk}/n) \cdot \bar{z}_{jk}^n$. By the continuity, and hence boundedness, of S, terms in this sum will be asymptotically negligible unless $n_{jk} \to \infty$, i.e., s_{jk} is infinite. Terms with $|j - k| \leq 1$ will be zero. We must therefore investigate the terms with $|j - k| > 1$ and infinite s_{jk}.

First suppose $j < k - 1$. Then $\bar{z}_{jk}^n = S(\bar{a}_{jk}^n, l_k) - S(\bar{a}_{jk}^n, u_j)$ with $u_j < l_k$. Let $\bar{p}_{jk}^n$ denote the average of the (p_i) for i in s_{jk}^n, and $w_{jk}^n = S(\bar{p}_{jk}^n, l_k) - S(\bar{p}_{jk}^n, u_j)$. Since **p** is computably calibrated and s_{jk} is computable and infinite, we can find T_{jk} such that, for $n > T_{jk}$, $|\bar{a}_{jk}^n - \bar{p}_{jk}^n| < \delta$, and then $|\bar{z}_{jk}^n - w_{jk}^n| < 2\varepsilon$. But, for $i \in s_{jk}$, $p_i \leq u_j$, whence $\bar{p}_{jk}^n \leq u_j < l_k$. Since S is a proper scoring rule, it follows (Savage, 1971, end of Section 4) that $w_{jk}^n \geq 0$. Thus, for $n > T_{jk}$, $\bar{z}_{jk}^n \geq -2\varepsilon$. Since a similar argument holds when $j > k + 1$, it follows that we can find T

such that $n > T \Rightarrow \bar{z}_{jk}^n \geq -2\varepsilon$, all j, k. Thus $\bar{z}_n \geq -2\varepsilon$, whence $\bar{y}_n \geq -4\varepsilon$ ($n > T$). □

9. The information base. So far we have supposed that the only data that can be used in formulating forecasts are the outcomes of previously forecast events. However, this is far too restrictive as a model for most practical forecasting problems, where the forecaster will usually have further relevant information. Thus the meterologist forecasting rain might have current and historical data on temperature, cloud formation, wind speed, etc.; the econometrician forecasting inflation might wish to take into account related series such as balance of payments and unemployment; and the utility company forecasting electricity demand will need to take account of external temperature. Our analysis can be extended to account for such expanded information.

Let $\mathscr{B}_n$ be the σ-field of events whose truth or falsity will be determined and supposed known to the forecaster at time n. We call $\mathscr{B} = (\mathscr{B}_n)$ the *information base*. Throughout this section we shall suppose $\mathscr{B}_n \subseteq \mathscr{B}_{n+1}$ (forecasters never forget) and $A_n \in \mathscr{B}_n$ (past outcome data are always available). The probability forecast p_{n+1} for A_{n+1} at time n is considered to be the realisation of a $\mathscr{B}_n$-measurable quantity $P_{n+1} \in [0, 1]$. A sequence $(P_1, P_2, \dots)$ of such quantities forms a *$\mathscr{B}$-forecasting system*. Such a system has the flexibility to take account of all the data available when forecasts are issued. Clearly, any probability distribution Π over $\mathscr{B}_\infty = \lim_{n\to\infty} \mathscr{B}_n$ determines such a system, with $P_{n+1} = \Pi(A_{n+1}|\mathscr{B}_n)$, and any such system is consistent in this way with such a distribution Π (and, generally, with many such).

In order to investigate the empirical validity of a $\mathscr{B}$-forecasting system F for **A**, we must take into account, not only the outcomes **a** of **A**, but also the data used in constructing the forecasts, since we must check that these data have been fully and appropriately utilised. We can express such data, specifying, as well as **a**, the truth or falsity of every $B_n \in \mathscr{B}_n$, for all n, as an elementary event $\beta \in \mathscr{B}_\infty$. Thus our validation check should involve some sort of comparison of F with β, the true "state of the world."

The metacriteria of Section 4 may be modified to apply to this new problem, on replacing references to the realised outcome sequence **a**, and the σ-field $\mathscr{A}_\infty$, by respective references, instead, to the realised elementary event β, and the σ-field $\mathscr{B}_\infty$. (We may also ignore the second sentence of M1, as no longer relevant.) With these changes, the same arguments for these metacriteria may be made as before. In particular, Theorem 4.1 continues to hold, with $\mathscr{A}_n$ and $\mathscr{A}_\infty$ replaced by $\mathscr{B}_n$ and $\mathscr{B}_\infty$, respectively, and its corollary, with P_{n+1}^i defined as $\Pi^i(A_{n+1}|B_n)$.

Once again we shall use calibration as the basis of our validity test, and once again we imagine an adversary who attempts to select, if possible, a subsequence of events with respect to which calibration fails. However, this time, in order to challenge the claim that the forecasts take full account of all the available information, the adversary is allowed to base his decision as to whether any event is to be included in his subsequence on the full information supposed available to the forecasting system when producing its probability forecast for that event.

Such an adversary may be represented by a *$\mathscr{B}$-selection rule* $\delta = (d_1, d_2, \dots)$, where d_{n+1} is a $\mathscr{B}_n$-measurable quantity, with possible values 0 or 1, such that the $(n+1)$th event is to be included in the selected subsequence if and only if $d_{n+1} = 1$. Our validity criterion will, as before, demand calibration with respect to all the subsequences selected by a suitable collection of such $\mathscr{B}$-selection rules. This will satisfy the modified metacriteria; in particular, the theorem of Dawid (1982a) still applies to show that, with Π-probability 1, the forecasts produced by a distribution Π over $\mathscr{B}_\infty$, with respect to an expanded information base $\mathscr{B}$, will be completely calibrated for the outcome sequence with respect to any countable collection of $\mathscr{B}$-selection rules.

To make further progress we must, for reasons already rehearsed, again impose suitable constraints of computability on both the $\mathscr{B}$-forecasting systems and the $\mathscr{B}$-selection rules admitted. The following extension of Definition 6.1 will be adequate for our purposes.

DEFINITION 9.1. Let $\mathscr{C} = (\mathscr{C}_1, \mathscr{C}_2, \dots)$ be an information base such that each $\mathscr{C}_n$ is generated by an (at most) countable partition $\Gamma_n = (C_{nj}: 1 \le j < J_n \le \infty)$. Then we call $\mathscr{C}$ *simple*. In this case, a $\mathscr{C}$-forecasting system F is called *simply $\mathscr{C}$-computable* if its forecasts for any data in $\mathscr{C}$ are all of the form $a2^{-b}$, with a and b integers, and there exists a Turing machine, which, when fed with an input tape containing (a finite encoding of) any two integers $n \ge 0$ and $j < J_n$, stops after a finite number of operations, having output the binary expansion of the probability forecast produced by F for A_{n+1} given the event $C_{nj} \in \mathscr{C}_n$.

A forecast sequence $\mathbf{p}$ is *$\mathscr{B}$-computable* for a realised elementary event $\beta \in \mathscr{B}_\infty$ if there exists a simple information base $\mathscr{C}$, with $\mathscr{C}_n \subseteq \mathscr{B}_n$ (all n), and a simply $\mathscr{C}$-computable forecasting system F, such that $p_{n+1} - F(c_n) \to 0$ $(n \to \infty)$, where c_n is the event of the partition Γ_n which obtains under β.

A subsequence s is *$\mathscr{B}$-computable* if it is so when considered as the sequence of forecasts σ, where $\sigma_n = 1$ if $n \in s$, $\sigma_n = 0$, otherwise.

A forecast sequence $\mathbf{p}$ is *$\mathscr{B}$-computably calibrated* for $\beta \in \mathscr{B}_\infty$ if it is calibrated for β with respect to every $\mathscr{B}$-computable subsequence.

THEOREM 9.1. *Let $\mathbf{p}^1, \mathbf{p}^2$ be $\mathscr{B}$-computable forecast sequences for $\beta \in \mathscr{B}_\infty$, each $\mathscr{B}$-computably calibrated for β. Then $p_n^1 - p_n^2 \to 0$ $(n \to \infty)$.*

The proof parallels very closely that of Theorem 7.1, and is therefore omitted. The theorem again asserts that, when judged by the validity criterion of calibration, all forecasting systems taking full and valid account of the same information base must make asymptotically indistinguishable probability forecasts.

We can similarly show, parallelling Theorem 8.1, that if $\mathbf{p}$ and $\mathbf{q}$ are $\mathscr{B}$-computable forecast sequences for β, and $\mathbf{p}$ is $\mathscr{B}$-computably calibrated for β, then $\mathbf{p}$ is S-superior to $\mathbf{q}$ with respect to $\mathbf{a}$ for any continuous proper scoring rule S. In general, if $\mathscr{B} \subset \mathscr{D}$, valid $\mathscr{D}$-computable forecasts will be strictly S-superior to valid $\mathscr{B}$-computable forecasts.

10. Logical probability. We have seen that the empirical validity of a sequence of forecasts depends, not only on the outcomes, but also on the information supposedly utilised in making those forecasts. Clearly, the more information the better. If $\mathscr{B}$ and $\mathscr{D}$ are information bases with $\mathscr{B} \subseteq \mathscr{D}$ (i.e., $\mathscr{B}_n \subseteq \mathscr{D}_n$, all n), then valid $\mathscr{D}$-computable forecasts will be $\mathscr{B}$-computably calibrated. However, such forecasts will not, in general, be $\mathscr{B}$-computable, and so not available when given only the information base $\mathscr{B}$. The $\mathscr{D}$-computable forecasts have the flexibility to respond more sensitively to relevant information; on the other hand, in order to be valid, they must respond appropriately by being calibrated with respect to a larger collection of subsequences than required for forecasts based on $\mathscr{B}$ alone. This stronger constraint limits the additional freedom available, to the extent of implying asymptotic uniqueness of valid $\mathscr{D}$-computable forecasts.

Since any information base $\mathscr{B}$ thus generates essentially unique valid forecasts with respect to $\mathscr{B}$, we can think of these forecast probabilities as expressing an objective, quasi-logical relationship between the information utilised and the outcomes. In effect, they provide a measure of "partial implication", i.e., the strength with which it is reasonable to assert that the forecast events will occur, on the (generally inconclusive) evidence of the data gathered. Thus they partake of some of the flavour of the logical probability concepts of Keynes (1921) and Carnap (1950), while remaining firmly tied to the specific empirical data which Nature chooses to produce.

Consider, for instance, a "deterministic" problem, where, for *some* suitably complete information base $\mathscr{D}$, a prediction with certainty is possible. [In particular, for any valid $\mathscr{D}$-computable probability forecasts $\mathbf{q}$, $q_n(1 - q_n) \to 0$ as $n \to \infty$.] In the context of forecasting rain, $\mathscr{D}$ might specify, to great precision, the positions and momenta of all particles in the atmosphere and oceans, and this might be enough to determine the next day's weather with certainty. Even repeated coin flips could be regarded as deterministic in this sense, with $\mathscr{D}$ specifying sufficient details of the angular momentum imparted in tossing, the physical layout of the table, etc. In practice, however, this "deterministic information base" will frequently not be available, but, instead, a much reduced information base $\mathscr{B} \subseteq \mathscr{D}$, no longer sufficient to allow pure computation of the future. Then valid $\mathscr{B}$-computable forecasts $\mathbf{p}$ will be asymptotically nondegenerate, with $p_n(1 - p_n) \not\to 0$. Thus we see that a nondegenerate probability need not necessarily be interpreted as measuring any intrinsic stochasticity in Nature. Rather, it can be considered as a price that must be paid for attempting to forecast on the basis of incomplete information.

11. Restricted information. Instead of considering an expanded information base, as in Section 9, we can generalise in the opposite direction, by limiting the information that may be used in making a forecast for A_{n+1} to be strictly less than the outcomes $\mathbf{a}^{(n)}$ of all past forecast events. For example, we might retain only the outcome a_n of the previous event A_n, thus restricting the forecasting system to have the form of a computable function $p_{n+1} = F(n, a_n)$. A further restriction might even exclude dependence on n from this formula, so that

$p_{n+1} = F(a_n)$, yielding $p_{n+1} = \lambda$, say, if $a_n = 1$, $p_{n+1} = \mu$ if $a_n = 0$. Such forecasts would be fully appropriate under a simple stationary Markov chain model for **A**, but not so otherwise. Nevertheless, if the forecasts are constrained in such a way, we can still ask whether they are in the best possible correspondence with the data, subject to the constraints.

Once again, calibration can be used to formalise such questions. We shall require calibration with respect to a subsequence selected by an adversary who is allowed to construct his selection rule on exactly the same basis as the forecasts under test. Thus, for forecasts restricted to the form $p_{n+1} = F(a_n)$, the only admissible nontrivial selection rules would pick out *either* the subsequence s_1 of all events following a success, *or* the subsequence s_0 of all events following a failure. The forecasts with $F(1) = \lambda$, $F(0) = \mu$ will thus be acceptable, subject to the restriction on their form, if and only if the limiting relative frequency of successes is λ in s_1 and μ in s_0.

The calibration criterion remains an intuitively appealing one in this context, although the arguments in its favour are less strong. In particular, M3 need not hold. Consider, for example, the severest possible restriction, $p_{n+1} =$ constant (taking no account of n and past data). The only admissible selection rules for this restriction yield either the null sequence or the complete sequence, so that the forecasts $p_{n+1} = \pi$ are acceptable if and only if π is the limiting relative frequency, r say, of success in the whole sequence.

A distribution Π over $\mathscr{A}_\infty$ may be considered as giving rise to such a forecasting system if the marginal probabilities $\Pi(A_{n+1})$ (which take no account of past data) are all equal (and thus take no account of n), and we could then take $p_{n+1} = \Pi(A_{n+1}) =$ constant. Consider, however, the exchangeable distribution Π obtained by mixing the Bernoulli trials model with respect to a uniform distribution for its parameter p. Then $p_{n+1} \equiv \frac{1}{2}$ and thus calibration holds if and only if $r = \frac{1}{2}$. However, with Π-probability 1, $r = p \neq \frac{1}{2}$. Thus M3 fails.

If, nonetheless, the restricted calibration criterion is applied, our principal result, Theorem 7.1, extends to this setup. The result will hold so long as, whenever F^1 and F^2 are admissible forecasting systems, δ_K, defined as in the proof of the theorem, is an admissible selection rule.

12. Prognostic systems. We now introduce another variation on the theme of empirically valid probabilties. We suppose that we have a large (conceptually infinite) ordered population of individuals $1, 2, \ldots$. Attached to each individual n is some given background information $\mathbf{x}_n$ and also an uncertain event A_n. We seek a rule π that attaches to each individual n a probability $p_n = \pi(\mathbf{x}_n)$, to be interpreted as the probability that A_n will occur, based on data $\mathbf{x}_n$.

Examples of such a setup include:

(a) Medical prognosis (or diagnosis), in which $\mathbf{x}_n$ consists of various items of patient n's clinical history, medical symptoms, etc. and A_n denotes the event that patient n will recover (or, has a particular disease);

(b) Insurance portfolios, in which $\mathbf{x}_n$ determines various actuarial rating factors and other properties of car driver n and A_n is the event that he or she will have an accident within a given year;

(c) Criminal trials, with $\mathbf{x}_n$ representing the evidence before the jury concerning defendant n and A_n the event of guilt.

It is common in such cases to build a statistical model relating p_n to $\mathbf{x}_n$. See for example, Titterington et al. (1981) for case (a), and Du Mouchel (1983) for case (b). Case (c) has not, to my knowledge, been subjected to such a treatment. However, our analysis will be general enough to allow such cases, even though the very nature of the information $\mathbf{x}$ may vary from one individual to another, and no two different individuals need have identical values for $\mathbf{x}$.

For concreteness, we shall refer to a rule π, specifying p_n completely as a function of $\mathbf{x}_n$, as a *prognostic system*. A standard statistical investigation might proceed by postulating a parametric or nonparametric family of such prognostic systems for a given setup, assuming that one of these is true, collecting some data on $(\mathbf{x}_n, A_n)$ for various individuals n, and attempting to use this to make inferences about the true underlying prognostic system.

We shall concern ourselves here with the meaning of the assumption that there exists a "true" prognostic system to be discovered by sufficiently extensive data analysis. There is no real problem if it can be assumed that, for any possible value $\mathbf{x}$ of the background information, the conceptual infinite population contains an infinite subset of individuals n having $\mathbf{x}_n = \mathbf{x}$; the probability p_n appropriate to any such individual would be taken to be the limiting proportion of individuals, in the subset, for which the associated event A occurs. This could be discovered, at least asymptotically, given enough data.

More interesting, however, is the contrary case. Is there any valid objective meaning to be attached to the probabilities associated with different individuals, when we cannot group them into large homogeneous subsets?

Suppose,then, we have a prognostic system π. What criteria should we demand it satisfy in order to be acceptable, in the light of sufficient data? As in earlier sections, our approach to this will be essentially infinitary, so that we shall suppose complete information available on $(\mathbf{x}_n, A_n)$ for all individuals. The important practical problem of the acceptability of π on the basis of finite data will not be tackled. However, our conclusion will be important as a philosophical justification for the type of statistical exercise commonly undertaken.

Although the background information may be so detailed as to give no way of decomposing the population into infinite homogeneous subsets *a priori*, the assumed prognostic system π itself imposes on the population a suitable decomposition. We may take, for example, all those individuals n for which $|p_n - 0.4| < 0.05$. If π is to be empirically meaningful, we should expect that in this subset, if infinite, the limiting relative frequency of occurrence of the associated events (a_n) should be 0.4 (to one decimal place). Moreover, since π claims not to recognise any noticeable differences in prognosis between the individuals in this subset, the same limiting proportion 0.4 should hold for a subsubset, chosen on the basis of the same background information $(\mathbf{x}_n)$ available to π. The parallels with our previous calibration criteria will be clear.

To be precise, we now consider only computable prognostic systems, wherein $\pi(x)$ is constrained to be a computable function of $\mathbf{x}$ (supposed to be approximately encodable as a finite string). We shall test these with computable selection

rules, i.e., those computable functions of $\mathbf{x}$ having values 1 (inclusion) or 0 (exclusion) only. And we shall require that π's sequence $\mathbf{p}$ of probabilities should be calibrated for the outcomes $\mathbf{a}$ with respect to all such computable selection rules, in which case we may again call π *computably calibrated.* This seems to be a reasonable requirement that π is getting overall proportion correct, and at the same time is making fullest possible use of the available background information in determining its prognostic probabilities.

In exactly the same way as in Theorem 7.1, we can now show the essential uniqueness of a computable and computably calibrated set of prognostic probabilities. If (p_n^1) and (p_n^2) both have this structure, then $p_n^1 - p_n^2 \to 0$ as $n \to \infty$. There thus exists (assuming calibrability, at least) an essentially unique set of objective prognostic probabilities based on the available data $\mathbf{x}$. An attempt to make inferences about these objective probabilities is therefore justified to the extent that it is a hunt for something which does, at least, have a unique existence, at any rate asymptotically. Note, however, that, just as before, the objective prognostic probabilities will depend on the extent and form of the information $\mathbf{x}$ used for prognosis.

13. Concluding remarks

13.1. Subjectivist implications. While the theory presented here can be considered as an extension of von Mises' frequency theory of probability, it in fact arose from an attempt to provide an empirical assessment of the sequence of forecasts produced by a coherent subjectivist forecaster. Thus p_{n+1} would represent the forecaster's strength of belief in A_{n+1} in the light of his past data.

The theory of de Finetti (1975) would put no constraints on subjective probability assignments other than that they be consistent with some probability distribution Π. In our case, this imposes no constraint at all on the (p_n), apart from the trivial requirement $0 \leq p_n \leq 1$. From this extreme viewpoint it would therefore seem that any set of forecasts is as good as any other. But this is to ignore the evident fact that some forecasters are more successful than others, and that some measure of empirical success is required as an external validation of subjectivist forecasts. As shown in Dawid (1982a), the calibration criterion we have chosen can itself be regarded as a consequence of a very natural requirement: A subjective distribution is discredited if a prespecified event to which it gave probability 1 fails to materialise.

It is a perhaps of surprising consequence of this validity criterion that it imposes asymptotic uniqueness on subjective probability forecasts. Any forecaster whose forecasts are not, ultimately, indistinguishable from the objective ones will eventually be discredited. This raises difficulties for the forecaster who cannot guarantee that he will produce objective forecasts. It also means that the scope for subjective disagreement between different forecasters is virtually eliminated, if they all wish to stay in touch with reality. In defence of the subjectivist Bayesian position, however, it should be pointed out that no other method of forecasting can be guaranteed to do any better (Dawid, 1985b).

13.2. Martingale extensions. If a forecaster assigns probability forecasts (p_n) to (A_n), he should be willing to accept bets on the (A_n) at odds determined by his probabilities. Since his forecasts are intended to take full account of his information base $\mathscr{B}$, his betting opponent should be allowed to exploit any departures he feels the forecaster is making from full use of this information, by allowing the size c_{n+1} of his bet on A_{n+1} to vary accordingly. The opponent's gain from such a bet will be $c_{n+1}(1 - p_{n+1})$ if A_{n+1} occurs, $-c_{n+1}p_{n+1}$ if not.

A $\mathscr{B}$-computable sequence of real-valued terms, $\mathbf{c} = (c_1, c_2, \ldots)$, may be termed a *betting strategy*. The accumulated fortune, by time n, of an opponent using strategy $\mathbf{c}$, for outcome sequence $\mathbf{a}$, will be $f_n = f_0 + \sum_{r=1}^{n} c_r(a_r - p_r)$.

The calibration criterion only covers the case where $\mathbf{c}$ is a selection rule, so that each $c_n = 0$ or 1. The opponent can choose whether or not to bet at any time but not the size of the bet. And Definition 5.1 requires that, for such a case, the opponent's total fortune f_n grows infinitely more slowly than the cumulative size $t_n = \sum_{r=1}^{n} c_r$ of the bets.

We can modify the calibration criterion by allowing the opponent an arbitrary betting strategy, and imposing suitable restrictions on the behaviour of f_n. Note that, under a probability model Π for which $p_r = \Pi(A_r|\mathscr{B}_{r-1})$, (f_n) forms a martingale with respect to the information base $\mathscr{B}$. Thus we can test the validity of this forecasting system by requiring that (f_n) should "look like" a martingale realisation. One way of formalising this, generalizing ideas of Ville (1939) and Schnorr (1971) for the Bernoulli case, is as follows. Consider an opponent who starts off with unit capital $f_0 = 1$. At any time he may choose c_n, as a function of past data, subject to the restriction that he must always have enough capital to meet his debt if he loses. In this case we may call the bet sequence $\mathbf{c}$ *allowable*, and the fortune sequence $\mathbf{f} = (f_n)$ *strongly nonnegative*. If the forecasts $\mathbf{p}$ made by Π are "correct", $\mathbf{f}$ will be a realisation of a nonnegative martingale, of unit mean. Such a martingale must be bounded above, with Π-probability 1. We can therefore impose, as a new validity criterion, the requirement that, for any allowable $\mathscr{B}$-computable bet sequence $\mathbf{c}$, the associated fortune sequence $\mathbf{f}$ is bounded above. This may be shown to be essentially the same as the requirements of Howard (1975) and Martin-Løf (1966). This martingale criterion says that, when betting at "correct" odds, it is impossible to make an unlimited fortune out of a finite initial capital. As a basis for a theory of probability, it has much in common with (and is as soundly established in practice as) the principle of the impossibility of a perpetual motion machine as a basis for physics.

The above martingale criterion may be shown to satisfy M1–M4 of Section 4, extended where necessary as in Section 9. It appears to be strictly stronger than the calibration criterion, just as Ville's development for the Bernoulli case strengthens that of von Mises. Further study of this criterion would appear promising.

13.3. Some analogies. The calibration criterion, and especially its martingale extension above, is somewhat analogous to de Finetti's coherence criterion.

Suppose a subjectivist assesses probabilities (p_i) (simultaneously, not sequentially) for a set of events (A_i). If these truly represent his beliefs, then he should be willing to accept as fair any combination of bets, of sizes (c_i), where his loss from the ith bet will be $c_i(A_i - p_i)$. The (p_i) are said to be *coherent* if, no matter what the sizes (c_i), be they positive or negative, of the bets selected by his opponent, the forecaster's total gain $-\Sigma_i c_i(a_i - p_i)$ will be nonnegative for at least one logically possible set of outcomes (a_i) of the (A_i). This holds if and only if there exists a probability distribution Π such that $p_i = \Pi(A_i)$.

The above martingale criterion extends this idea to the sequential case, but with constraints on the forecaster's total loss for the state of the world which actually obtains. It thus represents a way of taking standard Bayesian arguments based on the self-consistency of subjectivist beliefs and extending these to take account of the connexion between those beliefs and the empirical world.

Another idea closely analogous to both calibration and coherence is that of the *relevant subset* due originally to Fisher (1956) and developed by Buehler (1959). Let X have distribution P_θ governed by the parameter θ. Suppose that some method of inference produces, for each value x of X, an interval $I(x)$ of θ values, together with a "confidence coefficient" γ relating to the possibility that $\theta \in I(x)$. An adversary is allowed to select a subset S of $\mathbf{x}$ values, for which, say, he regards γ as an overestimate. He is successful, and the method of inference is thereby discredited, if he can do so in such a way that $P_\theta(\theta \in I(X) | X \in S) < \gamma$ for all θ. Extensions of this idea (Pierce, 1973) allow the opponent to lay bets against "$\theta \in I(x)$," the size $s(x)$ of the bet being allowed to depend on x; the original criterion is recovered if $s(x) = 0$ or 1 only.

13.4. Data analysis. How are we to assess calibration in the real world, given a finite sequence of probability forecasts and their associated outcomes? This is an extension of the problem of defining the "randomness" of a finite sequence: see e.g., Fine (1973, Chapter V). One cannot reasonably expect finite calibration, even approximately, with respect to all possible computable selection rules, for one of these would select, "by accident," just those events which in fact occurred. This is the well-known problem that any finite set of data will exhibit peculiarities and departures from expectations, but these may well be noise, not signal, and, if so, need not be taken seriously. One possible suggestion is that we might choose some collection of computable selection rules, ordered in some reasonable way (for example, in terms of some measure of their complexity, as considered by Kolmogorov, 1963), yielding a sequence $\delta_1, \delta_2, \ldots$. (Note that such a collection must exclude some computable selection rules, since it is impossible computably to order all of them—see Note 7.1). We further choose a function $k(n)$, tending to infinity more slowly than n (perhaps, following Kolmogorov, $k(n) \sim \alpha \log n$ would prove suitable). For a data sequence of length n, we could assess the departure from calibration with respect to $\delta_1, \delta_2, \ldots, \delta_{k(n)}$ only, for example by means of suitable significance tests, and regard the sequence as acceptable if it passed all these tests. One might well take δ_1 to select *all* events (i.e., first

investigate only the overall proportion), with $k(n) = 0$ for $n \leq n_1$, $k(n) = 1$ for $n_1 < n \leq n_2$ where $1 \ll n_1 \ll n_2$. Thereafter, different collections and orderings of the δs will be appropriate, depending on the kind of plausible and interesting departures from calibration which it is desired to pick up quickly. Further work is clearly needed to make a practicable method, necessarily somewhat *ad hoc*, out of these suggestions.

13.5. A case for empirical probability? This paper has set out a theory which might be regarded as justifying a concept of probability based on empirical correspondence with the real world. As such it might seem to give some comfort to statisticians of the frequentist school, and to discomfort the subjectivist. Closer attention to the nature of the justification, however, might reverse these conclusions. Empirically valid probabilities exist, essentially, only at infinity—no finite collection of probability forecasts can be declared invalid. (This conclusion holds, as argued in Section 4 above, for any criterion of empirical validity which satisfies the metacriteria of that section.) Furthermore, empirical probabilities cannot be calculated, in general (see Note 7.1)—success or failure at specifying valid probabilities is a matter of luck. With conclusions like this, our investigation of theories of empirical probability might be regarded as yielding a *counter-example* to the idea that it is a meaningful and useful concept. This position has been forcibly argued by Schervish (1983).

Acknowledgment. This paper incorporates and supersedes that of Dawid (1982b), which formed the basis of the author's Special Invited Paper presented at the Eastern Regional Meeting of the Institute of Mathematical Statistics in Nashville, Tennessee in March 1983. Some of the additional material was presented at the Annual Conference of the British Society for the Philosophy of Science, Brighton in September 1983. The seeds of the ideas developed here were originally planted in discussions with Don Rubin and Mike Titterington in April 1982, when the three of us were Invited Research Fellows at the Mathematical Research Center, University of Wisconsin, Madison, an environment which provided a stimulating climate for their germination. Their further growth has been nurtured by helpful comments from many people, including especially Carl Morris, Mark Schervish, Chris Wallace, John Howard, and Trevor Fenner.

REFERENCES

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

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with Discussion). *J. Roy. Statist. Soc. A* **143** 383–430.

BUEHLER, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30** 845–863.

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

CHURCH, A. (1940). On the concept of a random sequence. *Bull. Amer. Math. Soc.* **46** 130–135.

CURTISS, J.H. (1968). An elementary mathematical model for the interpretation of precipitation probability forecasts. *J. Appl. Meteor.* **7** 3–17.

DAWID, A. P. (1982a). The well-calibrated Bayesian (with Discussion). *J. Amer. Statist. Assoc.* **77** 605–613.

DAWID, A. P. (1982b). Objective probability forecasts. Research Report 14, Department of Statistical Science, University College London.

DAWID, A. P. (1984). Statistical theory. The prequential approach (with Discussion). *J. Roy. Statist. Soc. A* **147** 278–292.

DAWID, A. P. (1985a). Probability forecasting. In *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson, and C. B. Read, eds.). Wiley-Interscience. To appear.

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

DE FINETTI, B. (1975). *Theory of Probability* (English translation). Two volumes. Wiley, New York.

DU MOUCHEL, W. H. (1983). The 1982 Massachusetts automobile insurance classification scheme. *The Statistician* **32** 69–81.

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70** 193–242.

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

FISHER, R. A. (1956). On a test of significance in Pearson's *Biometrika Tables* (No. 11). *J. Roy. Statist. Soc. B* **18** 56–60.

HOWARD, J. V. (1975). Computable explanations. *Z. Math. Logik Grundlag. Math.* **21** 215–224.

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

KNUTH, D. E. (1969). *The Art of Computer Programming: 2. Seminumerical Algorithms.* Addison-Wesley, Reading, Mass.

KOLMOGOROV, A. N. (1963). On tables of random numbers. *Sankhyā Ser. A* **25** 369–376.

KUHN, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

LICHTENSTEIN, S., FISCHHOFF, B. and PHILLIPS, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In *Judgment Under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic, and A. Tversky, eds.). Cambridge University Press.

MARTIN-LØF, P. (1966). The definition of random sequences. *Inform. and Control* **9** 602–619.

MARTIN-LØF, P. (1969). The literature on von Mises' Kollektivs revisited. *Theoria* **35** 12–37.

VON MISES, R. (1936). *Wahrscheinlichkeit, Statistik und Wahrheit.* Springer, Vienna. English translation: *Probability, Statistics and Truth* (1957). George Allen and Unwin, London.

OAKES, D. (1985). Self-calibrating priors do not exist. *J. Amer. Statist. Assoc.* **80** 339.

PIERCE, D. A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1** 241–250.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

SCHERVISH, M. J. (1983). There are no objective probability forecasts. Technical Report 277, Department of Statistics, Carnegie-Mellon University.

SCHERVISH, M. J. (1985). Comment on "Self-calibrating priors do not exist" by David Oakes. *J. Amer. Statist. Assoc.* **80** 341–342.

SCHNORR, C. P. (1971). A unified approach to the definition of random sequences. *Math. Systems Theory* **5** 246–258.

SUDBURY, A. W. (1973). Could there exist a world which obeyed no scientific laws? *British J. Philos. Sci.* **24** 39–40.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. D., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. and GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients (with Discussion). *J. Roy. Statist. Soc. A* **144** 145–174.

TURING, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.* (2) **42** 230–265.

VILLE, J. (1939). *Etude Critique de la Notion de Collectif.* Gauthier-Villars, Paris.

WALD, A. (1937). Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung. *Ergebnisse eines mathematischen Kolloquiums* **8** 38–72.

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
LONDON WC1E 6BT
ENGLAND

DISCUSSION

MARK J. SCHERVISH

Carnegie-Mellon University

I wish to thank Professor Dawid for providing such a thought-provoking paper to discuss. He has raised an interesting question in his paper, namely, "Do objective probabilities for events exist, relative to a given information base?" Professor Dawid suggests that the answer is yes, while this discussant believes that the answer is no.

1. Existence. Professor Dawid's main Theorems 7.1 and 9.1 prove the asymptotic closeness of computably calibrated computable forecasts. Their existence for any given forecasting problem is an open question. The purpose of this section is to cast doubt on their existence.

Whether or not there exists a single sequence of computably calibrated computable forecasts depends on exactly which sequence a actually occurs. Schervish (1985) has shown that there are uncountably many sequences a such that not a single computably calibrated computable forecasting system exists. That is, there are as many noncalibrable sequences as there are calibrable ones. The claim, which Professor Dawid makes, that the noncalibrable sequences are sparse in an intuitive sense, is an understandable outgrowth of the fact that, as statisticians, we view the world through the rose-colored glasses of computable forecasting systems. Hence, we see only calibrable sequences (with probability 1). But Nature is not (to my knowledge) hampered by the same computability restrictions as statisticians are. It follows, then, from the cardinality argument above that the most positive answer we can give to the question of the existence of objective probabilities is "Maybe they exist, maybe not." In Section 2 we will show that even such a weak positive answer is unwarranted.

Even if the sequence a is noncalibrable, there is no cause for alarm in the forecasting community. It may very well be the case that, for many forecasters, the majority of forecasts in any finite initial segment are still quite good. That is, most forecasts may still be close to the indicators of the forecast events.

2. Probabilities of events. Suppose that the sequence a which will occur will be calibrable. (Please, do not ask how we might know this.) What then are