The Well-Calibrated Bayesian

A. P. DAWID*

Suppose that a forecaster sequentially assigns probabilities to events. He is *well calibrated* if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent. We prove a theorem to the effect that a coherent Bayesian expects to be well calibrated, and consider its destructive implications for the theory of coherence.

KEY WORDS: Calibration; Coherence; Martingale; Probability forecasting; Subjectivism; Weather forecasting.

1. INTRODUCTION

Subjective probability forecasting is now well established among meteorologists, particularly in the United States (Murphy and Winkler 1977). Weather forecasters routinely make predictions such as "the precipitation probability for Denver today is 30 percent"; they have also experimented with credible interval temperature forecasts of the form "the probability is 75 percent that today's maximum temperature in Denver will be between 63° and 67°F." The probabilities quoted refer to the forecasters' subjective "degree of belief," given their information at the time of the forecast. This information may include the "objective forecast" output from a climatological analysis, or a computer forecasting system; however, no explicit modeling process need be involved in arriving at forecast probabilities.

Such probability forecasting fits neatly into the general Bayesian world-view as conceived by de Finetti (1975). The coherent subjectivist Bayesian can be shown to have a joint probability distribution over all conceivably observable quantities. Forecasting then is merely a matter of summarizing the conditional distribution of quantities still unobserved, given current information. In this article we shall, for definiteness, talk mainly in terms of weather forecasting, but it should be understood that the scope of the discussion is much wider, taking in all applications in which a subjectivist makes repeated probability forecasts. For added definiteness, and with the usual nonsexist understanding, we shall refer to the forecaster in the masculine.

Probability forecasts can be judged by several criteria (Murphy and Epstein 1967). In this article we concentrate exclusively on the criterion of *calibration* (sometimes termed *reliability*). Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long-run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's *empirical calibration curve*. If the curve is the diagonal, $p = \omega$, the forecaster may be termed (empirically) *well calibrated*. A parallel concept holds for credible interval forecasts: these are well calibrated if, for example, the long-run proportion of forecast 75 percent credible intervals that succeed in covering the actual value of the predicted quantity turns out to be 75 percent.

The calibration criterion has some similarity with the frequency definition of probability, but does not require a background of repeated trials under constant conditions. In particular, it is rarely appropriate to interpret a subjective probability forecast as an estimate of some underlying "objective" probability; it is usually better considered as an estimate of (the indicator of) the forecast event itself. Thus we do not have to concern ourselves with the "true" probability of rain on a given day. Roberts (1968) has attempted to interpret such a concept by supposing that one could select a subset of all days that could be regarded, at the time of forecast, as identical in all relevant respects, and consider the limiting relative frequency of rain on such days as the "true" probability for any one of them. However, it is doubtful whether such a selection is practically meaningful, or whether different forecasters would agree on it. The calibration approach avoids these difficulties.

Murphy and Winkler (1977) show that experienced weather forecasters are, on the whole, well calibrated. Although this is not by itself a sufficient condition for their forecasts to be "good" (it would hold, for example, for a forecaster who invariably gave the long-term relative frequency of rain as his precipitation probability), it has often been taken to be a minimal desirable property. Further empirical studies of calibration have been reported by Lichtenstein, Fischhoff, and Phillips (1977), who obtain some poorly calibrated responses.

A common suggestion (e.g., in Cox 1958) is that the probability statements of a poorly calibrated forecaster should be transformed before use. Thus, if a forecaster's empirical calibration curve at a quoted value of 30 percent has $p = 20$ percent, then the consumer of this forecast

* A. P. Dawid is Professor of Probability and Statistics, Department of Statistical Science, University College London, London WC1E 6BT, England. The author is grateful to the editor, the associate editor, a referee, and Professor M. Stone, for providing valuable comments and relevant references.

September 1982, Volume 77, Number 379
Theory and Methods Section

might perhaps best assume that the "true probability" of rain is 20 percent. We shall not consider here the general problem of how to use someone else's forecasts; for this, see for example Morris (1974), and Lindley, Tversky, and Brown (1979).

In this article, we investigate the forecaster's view of his own calibration, and show, in particular, that if he is coherent then he expects to be well calibrated. We also discuss the problems that this creates for the theory of coherence.

2. INDEPENDENCE AND FEEDBACK

Previous theoretical studies of the calibration property (Morris 1977; Harrison 1977) have mostly been concerned with the assessment of "seemingly unrelated" uncertain events or quantities: for example the height of the Eiffel Tower, the freezing point of mercury, the population of the USSR, and so on. It appears natural for the forecaster to regard such quantities as probabilistically independent. Supposing this, consider now the dilemma faced by such a forecaster who learns that of a very large number N of such forecasts that he has made 80 percent (say) exceed their assessed medians. This is an event whose subjective probability may be calculated from the binomial distribution with probability $\frac{1}{2}$, and will be vanishingly small; yet it occurred. In the light of such a conflict, it might seem appropriate to borrow from the logic of significance testing and reject the basis for the probability assessments. The moral that Harrison (1977) draws is that the naive approach, in which "seemingly unrelated" quantities were regarded as subjectively independent, is unacceptable; assessments for such quantities are, in fact, related merely by virtue of the fact that they are being made by the same assessor. Defining a *potentially miscalibrated individual* as one who is not sure whether his future subjective probability assignments will agree with observed frequency, Harrison goes so far as to conclude that "such a person will never perceive two events as (probabilistically) independent."

These investigations seem too specialized in contexts such as weather forecasting, where, for example, the criterion of "seeming unrelatedness" of precipitation for different days is clearly inapplicable. We shall follow a different path, taking advantage of the sequential nature of the weather forecaster's task. The forecaster does not operate by giving, on 31 December, his individual precipitation forecasts for every day of the coming year, and then retiring: each day he forecasts for tomorrow, drawing on his accumulated experience of all that has passed up to today, including, in particular, the outcomes of those of today's events for which he supplied forecasts yesterday. It is such *sequential forecasts with feedback* that will form our principal subject of study.

Our mathematical structure is as follows. Forecasts are made sequentially on days 0, 1, 2, . . . , each referring to events or quantities that will become known on the following day. We denote by $\mathcal{B}_i$ the totality of events known to the forecaster on day i; thus $\mathcal{B}_0 \subseteq \mathcal{B}_1 \subseteq \ldots$. The forecaster has an arbitrary subjective probability distribution Π defined over $\mathcal{B}_\infty = \bigvee_{i=0}^{\infty} \mathcal{B}_i$. The probability forecasts he makes on day i are for events or quantities in $\mathcal{B}_{i+1}$, and are calculated from his current conditional distribution $\Pi(\cdot \mid \mathcal{B}_i)$.

With this formulation, the problem of the badly calibrated forecaster is much more serious. For suppose X_i is a $\mathcal{B}_i$-measurable quantity ($i = 1, 2, \ldots$), for example the maximum temperature in Denver on day i, and let m_i be the median of the forecaster's distribution for X_i, as assessed on day $i - 1$. Let S_i denote the event "$X_i > m_i$". Then, by definition, $\Pi(S_i \mid \mathcal{B}_{i-1}) \equiv \frac{1}{2}$. Since $\mathcal{B}_{i-1}$ contains $(S_1, S_2, \ldots, S_{i-1})$, it readily follows that, according to Π, $\Pi(S_i) = \frac{1}{2}$ and the (S_i) are independent (Pratt 1962). Once again, we appear to have a conflict if, over many days, 80 percent, say, of the (X_i) exceed their assessed medians. However, the only assumption made above was that Π be coherent, so that Π obeys the laws of probability theory. That is, any coherent sequential forecaster must completely discount the possibility that he might be miscalibrated, however strong the evidence against him might be. In other words, in our sequential setup, Harrison's potentially miscalibrated individual cannot be coherent. We return to this point in Section 6.

3. A GENERAL CALIBRATION THEOREM

In this section we present a very general result that extends the above connections between coherence and calibration. Once again we suppose the forecasts are made sequentially according to a fixed probability distribution Π, but make no other assumptions.

For each day i we have an arbitrary associated event $S_i \in \mathcal{B}_i$, for example, the event of precipitation on day i. We denote the indicator of S_i by Y_i, and introduce $\tilde{Y}_i = \Pi(S_i \mid \mathcal{B}_{i-1}) = E(Y_i \mid \mathcal{B}_{i-1})$, the probability forecast of S_i on day $(i - 1)$.

One way of comparing forecasts with reality is to pick out some fairly arbitrary *test set* of days, and in it compare (a) the proportion p of days whose associated events in fact occur with (b) the average forecast probability π for those days. Formally, we introduce indicator variables $\xi_1, \xi_2, \ldots$, at choice, to denote the inclusion of any particular day i in the test set: $\xi_i = 1$ if day i is included, $\xi_i = 0$ otherwise.

We might choose the test set in advance, once and for all. However, it is a useful extension to allow the (ξ_i) themselves to be determined sequentially; thus the decision on inclusion or exclusion of day i need only be made on day $(i - 1)$, and then in an arbitrary way, in the light of knowledge available by day $(i - 1)$. Formally, ξ_i must be $\mathcal{B}_{i-1}$-measurable. Apart from this, no restriction whatsoever is placed on the selection of days into the test set. We call any such selection process *admissible*.

Let

$$\nu_k = \sum_{i=1}^{k} \xi_i, \; p_k = \nu_k^{-1} \sum_{i=1}^{k} \xi_i Y_i, \; \pi_k = \nu_k^{-1} \cdot \sum_{i=1}^{k} \xi_i \tilde{Y}_i.$$

That is, restricting attention to those days up to day k

selected for inclusion in the test set, ν_k is the number of such days, p_k the proportion for which the associated events in fact occur, and π_k the average forecast probability. Then we have the following result.

Theorem. Let the selection process (ξ_i) be admissible. With Π-probability one, if $\nu_k \to \infty$ then $p_k - \pi_k \to 0$.

The proof is given, with some extensions, in the Appendix. Note that the result could not be true in general if we were to allow ξ_i to depend on Y_i, for then we could force $p_k = 0$, for example.

4. APPLICATIONS

4.1. Empirical Calibration

Fix $\omega \in (0, 1)$, $\delta > 0$, and define $\xi_i = 1$ if and only if $| \bar{Y}_i - \omega | \leq \delta$. That is, our test set of days consists of just those for which the assessed probability of the associated event is suitably close to ω. This is admissible, since the condition determining ξ_i can be decided on day $(i - 1)$. For this choice, $| \pi_k - \omega | \leq \delta$. It thus follows from the Theorem that, with Π-probability one, assuming the selection condition is satisfied infinitely often, p_k will be close to ω for all sufficiently large k. That is to say, the coherent sequential forecaster believes that he will be empirically well calibrated.

An extension of the preceding result is obtained on choosing $\xi_i = 1$ when $| \bar{Y}_i - \omega | \leq \delta_i$, when δ_i is possibly allowed to depend on information up to day $(i - 1)$, and $\delta_i \to 0$. The conclusion then is that, with Π-probability one, if the sequence of selected days is infinite, $p_k \to \omega$.

4.2 Variable Event Calibration

At first sight, it seems that our Theorem does not cover the possibility that the event S_i is itself sequentially selected by the forecaster. For example, if X_i is the maximum temperature in Denver on day i, the forecaster may calculate his conditional distribution for X_i, given $\mathcal{B}_{i-1}$, and from it construct, say, some 75 percent credible interval $A_i \subseteq \mathbf{R}$, taking $S_i =$ "$X_i \in A_i$." But of course, even with this extension, the constructed S_i belongs to $\mathcal{B}_i$, so that the Theorem applies. In general, the only extra condition needed, satisfied in the above example, is that the determination of the variable event considered on day i shall be effected by day i.

In this example, $\bar{Y}_i \equiv .75$ by construction, whence $\pi_k \equiv .75$, and the Theorem entails the convergence of p_k to .75 with Π-probability one for any infinite admissible selection, and in particular for the whole sequence $(\xi_i \equiv 1)$. (Of course, this conclusion is already implicit in the argument of Section 2, which shows that the (S_i) behave, under Π, as Bernoulli trials with probability .75.) Thus the coherent forecaster expects his sequential credible interval forecasts to be well calibrated.

4.3 Model-Based Forecasts

Consider now the special case in which it can be agreed that the data arise from some "objective," unknown probability distribution P. Suppose our forecaster postulates a model "$P \in \mathcal{P}$," where $\mathcal{P} = \{P_\theta\}$, with $\theta \in \Theta$, a subset of $\mathbf{R}^k$. Suppose further that, for this model, θ is consistently estimable. His distribution Π is now completely specified by his prior distribution over Θ; we suppose that this is *full*, in other words has support Θ.

Under weak regularity conditions, if indeed $P \in \mathcal{P}$, say $P = P_{\theta_0}$, his posterior distribution for θ will (with P-probability one) converge to the one-point distribution at θ_0. This will be reflected in his probability forecasts, which will asymptotically approximate the "objective" probabilities under P_{θ_0}, and so be well calibrated with P-probability one, by our Theorem. Thus a full prior for a model that includes the true distribution P of the data will yield forecasts that will in fact, that is, under P, be (almost certainly) well calibrated. If the calibration property appears to fail, then the whole model is discredited.

As an example, suppose the forecaster postulates a Bernoulli model $\mathcal{P} = \{P_\theta\}$, where, according to P_θ, the $\{Y_i\}$ are independent with $P_\theta(Y_i = 1) = \theta$. For definiteness, take his full prior to be uniform on [0, 1]. If $\mathcal{B}_n$ only contains information on $(Y_1, \ldots, Y_n)$, his sequential probability forecast $\bar{Y}_{n+1}$ of Y_{n+1} is

$$P(Y_{n+1} = 1 \mid Y_1, \ldots, Y_n) = (r + 1)/(n + 2),$$

where r is the number of 1's in the first n Y's. If now r/n (and thus $\bar{Y}_n$) tends to a limit, λ say, as $n \to \infty$, then the forecasts will be empirically well calibrated (for only when $\omega \doteqdot \lambda$ do we get an infinite set of trials for which $| \bar{Y}_n - \omega | \leq \delta$ on which calibration can be tested and could fail; but this set will contain all trials beyond some point, and so yield $p_k \to \lambda \doteqdot \omega$). But, when $P \in \mathcal{P}$, r/n does, indeed, converge (with probability one).

Now in this case, the preceding simple empirical calibration criterion is a poor test of "$P \in \mathcal{P}$," for, even if $P \notin \mathcal{P}$, only for pathological P would r/n not converge almost surely to a limit. One could, instead, use the general Theorem, selecting say only those trials i for which $Y_{i-1} = 1$. When $r/n \to \lambda$, we again get $\pi_k \to \lambda$, so that these forecasts are well calibrated if and only if $p_k \to \lambda$, that is, the limiting relative frequency of 1's is the same following a 1 as overall. This occurs with probability one for the Bernoulli model, but would fail, for example, if the sequence $(Y_1, Y_2, \ldots)$ followed a general Markov Chain.

5. RECALIBRATION?

Suppose that you have made a large number of probability forecasts. On examining your empirical calibration curve, you find that it departs markedly from the diagonal. Can you learn about your own inadequacies as a forecaster from this, and use this knowledge to improve future assessments?

Various authors, for example Morris (1977) and Harrison (1977), have attempted to structure this problem along the following lines (a related, more complicated approach may be found in De Groot 1980). You model the various events to which you were initially willing to

assign some common probability ω (or some appropriate subset thereof) as *exchangeable*. It then appears to follow that, after observing a proportion $p \neq \omega$ of many such past events in fact occurring, the next such event (to which you wanted to give probability ω) should in fact be assigned a probability near p. That is, your ω is recalibrated to p.

While this seems very sensible, its coherence is suspect. How can you simultaneously assign two different probabilities to one event? The obvious answer is that they must be conditional on different information: ω is prior, and p posterior, to the calibration experience. Such a response, however, will not do when the initial probability assessments are sequential, since the calibration experience is then also prior to ω. As we have seen in Section 2, in this case all the events under consideration are judged independent. This is a degenerate case of exchangeability and does not allow for accumulated experience to alter probabilities. If you wish to recalibrate sequential forecasts, you are being incoherent.

Even if you do recalibrate, and eventually achieve a satisfactory empirical calibration curve, it does not follow that the property of the Theorem (π_k close to p_k) will hold for arbitrary admissible selections. Similar remarks apply to the forecaster who attempts to "cheat," by quoting probabilities that differ from his true assessments in an attempt to improve his apparent calibration performance (De Groot 1979). While this may be possible to a limited extent, it would not guarantee that π_k will be close to p_k in the general case.

6. COHERENCE AND CROMWELL'S RULE

Any application of the Theorem yields a statement of the form $\Pi(A) = 1$, where A expresses some property of perfect calibration for the distribution Π. In practice, however, it is rare for probability forecasts to be well calibrated (so far as can be judged from finite experience) and no realistic forecaster would believe too strongly in his own calibration performance. We have a paradox: an event can be distinguished (easily, and indeed in many ways) that is given subjective probability one and yet is not regarded as "morally certain." How can the theory of coherence, which is founded on assumptions of rationality, allow such an irrational conclusion? In order to answer this question, we must consider more deeply the foundations of the theory of coherence, and in particular, the interpretation of zero probabilities.

One approach to the theory of coherence is as follows (de Finetti 1964; Lehman 1955). Let A be an event, identified with its indicator. Your subjective probability of A is π if you would regard as fair a bet that returned you $c(A - \pi)$. Here c, related to the stake, is at choice, and may be positive or negative. (For realism, c should be small.)

If you now attach subjective probabilities (π) to various events (A), then you should regard as fair a combined bet that results from simultaneous fair bets, at arbitrary stakes, on a finite collection of these events. The return from such a combined bet would have the form $\sum_{i=1}^{n} c_i(A_i - \pi_i)$, where the (c_i) are arbitrary, and $\pi_i = \pi(A_i)$.

The *principle of coherence* requires that you do not regard as fair a bet whose return is certain to be negative, whatever the outcomes of the events involved.

It follows from this principle that (π) must be chosen to avoid the possibility that, for some choice of (A_i, c_i), $\sum_{i=1}^{n} c_i(A_i - \pi_i) < 0$ always. It may then be established, for example, that the (π) must lie in [0, 1], and obey the (finite) addition law of probability. An extension of this argument to called-off bets produces the multiplication law.

The above definition of coherence has been criticized as too weak by Shimony (1955) and Kemeny (1955). They prefer a *principle of strict coherence* (see Carnap 1971) that refuses to allow as fair a bet whose return is never positive, and sometimes negative. This possibility is allowed by our earlier (weak) principle of coherence, although the event of negative return must then be assigned zero probability. Strict coherence implies that no possible event can have probability zero, a property Carnap (1971) calls *regularity*. Lindley (1982) dubs this regularity requirement "Cromwell's rule."

Clearly, regularity cannot hold in continuous sample spaces, and the above principle of strict coherence becomes unworkable. Nevertheless, the weak principle still appears too weak; Buehler (1976), reflecting on his examples, opines "we have yet to arrive at a suitable theory of coherence for statistical models having arbitrary parameter spaces."

One possible position is as follows. In any event-field $\mathcal{A}$, there will be a class $\mathcal{I}$ of events that, while they may be logically possible, nevertheless can be regarded as "morally impossible" or "ignorable" (Dawid 1980): for example, the (idealized) event of a dart hitting an exactly specified point on the board. If we are prepared to countenance a bet that never wins, and loses sometimes, so long as the event of loss is ignorable, then we need only ensure that our subjective probability is positive for nonignorable events. We take this as the generalization of Cromwell's rule. I, at any rate, find such a principle compelling.

However, the property discussed at the start of this section implies that the typically nonignorable event of miscalibration must be assigned probability zero. While this does not contradict weak coherence, it is in conflict with the above appealing version of Cromwell's rule. (Although we have assumed countable, rather than finite, additivity in deriving our Theorem, I believe this does not alter the general conclusion if suitably interpreted.) As I am loth to accept a theory of coherence that does not contain some form of Cromwell's rule, my confidence in the universal applicability of the theory of coherence is shaken.

7. COHERENCE OR CALIBRATION?

The dilemma would be harmless if the forecaster were not an individual, but a constructed statistical system that outputs probabilities on being fed with appropriate data:

for example, a system for probabilistic medical diagnosis, tuned on a training set of patients, and applied to symptom information on new patients (Titterington et al. 1981). Any such system is applied only tentatively, while it seems to be working; as soon as it is clear that there is a conflict between its predictions and reality, such as clear evidence of miscalibration, the system will be modified or discarded. Because the system was never regarded as infallible, this causes no difficulty.

It seems to me that the subjectivist forecaster is obliged to treat his own subjective distribution Π in the same tentative manner as he would an external statistical forecasting system. If Π attaches probability zero to a nonignorable event, such as asymptotic miscalibration, and if this event happens, then Π must be treated with suspicion, and modified (e.g. by recalibration) or replaced. But such a process is intrinsically incoherent.

In practice, we should deal with an event such as poor calibration over a long historical sequence, suitably defined, and with its assigned near-zero probability. If the event is chosen in advance, at any rate, its occurrence must cast doubt on the distribution Π. This idea is close to classical hypothesis testing, and could have correspondingly many variants. Although I cannot perceive any clear logical principles that might govern its detailed application, I find its general message unavoidable. Box (1980) has put forward a similar view of scientific inference as the construction of successive Bayesian models of the world, each being subject to empirical test of the above kind, and replaced when it no longer seems to describe reality. A difficulty with this position is that one has no guarantee that the incoherent process suggested would perform any better (in calibration, say) than a coherent one.

The conflict between calibration and coherence could be avoided only by a distribution Π that was not even potentially miscalibrated. Such a distribution would have to take account of information in $\mathcal{B}_{i-1}$ about its calibration performance to date when forecasting for day i, as well as being fully coherent and representing the acceptable betting behavior of the forecaster. Considering the wide variety of admissible selections that may be used to test the calibration property, it seems doubtful, although not impossible, that such a coherent, self-calibrating distribution could exist.

APPENDIX: PROOF OF THEOREM

The proof is a slight variant of that of Theorem VII. 9.3. of Feller (1971). Let $\beta_i = v_i^{-1}$ if $v_i > 0$, $\beta_i = 0$ otherwise, and let $X_i = \beta_i\xi_i(Y_i - \tilde{Y}_i)$. Since β_i, ξ_i and $\tilde{Y}_i$ are $\mathcal{B}_{i-1}$-measurable, and $\tilde{Y}_i = E(Y_i \mid \mathcal{B}_{i-1})$, it follows that $E(X_i \mid \mathcal{B}_{i-1}) = 0$, so that, with $U_k = \sum_{i=1}^{k} X_i$, (U_k) is a martingale adapted to $(\mathcal{B}_k)$. Also,

$$E(X_i^2) = E[(\beta_i\xi_i)^2 \operatorname{var}(Y_i \mid \mathcal{B}_{i-1})] \le \tfrac{1}{4}E[(\beta_i\xi_i)^2],$$

so

$$E(U_k^2) = \sum_{i=1}^{k} E(X_i^2) \le \tfrac{1}{4}E\left[\sum_{i=1}^{k} (\beta_i\xi_i)^2\right].$$

Now for any realization of $(\xi_1, \xi_2, \xi_3, \ldots)$, the successive nonzero terms of the sequence $(\beta_1\xi_1)^2$, $(\beta_2\xi_2)^2$, . . . are $1, 1/2^2, 1/3^2, 1/4^2, \ldots$. Thus

$$\sum_{i=1}^{k} (\beta_i\xi_i)^2 \le \sum_{n=1}^{\infty} n^{-2} = \pi^2/6,$$

and so $E(U_k^2)$ is bounded above by $\pi^2/24$. By the martingale convergence theorem, the sequence $(U_k) = (\sum_{i=1}^{k} \beta_i \cdot \xi_i(Y_i - \tilde{Y}_i))$ converges with Π-probability one. From Kronecker's lemma (Lemma VII. 8.1 of Feller 1971, correcting a misprint), this convergence implies that

$$p_k - \pi_k = \beta_k \sum_{i=1}^{k} \xi_i(Y_i - \tilde{Y}_i) \to 0$$

so long as (β_k) tends monotonically to 0, which will hold when $v_k \to \infty$.

The Theorem and proof continue to hold for arbitrary random quantities (Y_i), not necessarily $0 - 1$, with $\tilde{Y}_i = E(Y_i \mid \mathcal{B}_{i-1})$, so long as $\operatorname{var}(Y_i \mid \mathcal{B}_{i-1})$ is uniformly bounded above; no doubt this condition could be relaxed. The identical argument in fact yields the more refined result that

$$g(v_k)^{-1} \sum_{i=1}^{k} \xi_i(Y_i - \tilde{Y}_i) \to 0$$

(Π – almost surely when $v_k \to \infty$) so long as $g(n)$ is eventually nondecreasing with $\sum_{n=1}^{\infty} g(n)^{-2} < \infty$. In particular, $p_k - \pi_k = 0(v_k^{-\alpha})$ for any $\alpha < \frac{1}{2}$.

[*Received November 1979. Revised January 1981.*]

REFERENCES

BUEHLER, R.J. (1976), "Coherent Preferences," *Annals of Statistics*, 4, 1051–1064.

BOX, G.E.P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness" (With Discussion), *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

CARNAP, R. (1971), "A Basic System of Inductive Logic (Part I)," *Studies in Subjective Logic and Probability*, eds. R. Carnap & R.C. Jeffrey, Berkeley: University of California Press, 35–165.

COX, D.R. (1958), "Two Further Applications of a Model for Binary Regression," *Biometrika*, 45, 562–565.

DAWID, A.P. (1980), "Conditional Independence for Statistical Operations," *Annals of Statistics*, 8, 598–617.

DE FINETTI, B. (1964), "Foresight: Its Logical Laws, Its Subjective Sources" (translated from French), in *Studies in Subjective Probability*, eds. H.E. Kyburg and H.E. Smokler, New York: John Wiley, 93–158.

——— (1975), *Theory of Probability*, New York: John Wiley.

DE GROOT, M.H. (1979), "Comments," *Journal of the Royal Statistical Society*, Ser. A, 142, 172–173.

——— (1980), "Improving Predictive Distributions," in *Bayesian Statistics*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Valencia: University Press, 385–395.

FELLER, W. (1971), *An Introduction to Probability Theory and its Applications* (vol. II, 2nd ed.), New York: John Wiley.

HARRISON, J.M. (1977), "Independence and Calibration in Decision Analysis," *Management Science*, 24, 320–328.

KEMENY, J.G. (1955), "Fair Bets and Inductive Probabilities," *Journal of Symbolic Logic*, 20, 263–273.

LEHMAN, R.S. (1955), "On Confirmation and Rational Betting," *The Journal of Symbolic Logic*, 20, 251–262.

LICHTENSTEIN, S., FISCHHOFF, B., and PHILLIPS, L.D. (1977), "Calibration of Probabilities: the State of the Art," in *Decision Making and Change in Human Affairs*, eds. H. Jungerman and G. de Zeeuw, Dordrecht: D. Reidel, 275–324.

LINDLEY, D.V. (1982), "The Bayesian Approach to Statistics," in

Some Recent Advances in Statistics, ed. J. Tiago de Olivera, London: Academic Press, 65–87.
LINDLEY, D.V., TVERSKY, A., and BROWN, R.V. (1979), "On the Reconciliation of Probability Assessments" (With Discussion), *Journal of the Royal Statistical Society*, Ser. A, 142, 146–180.
MORRIS, P.A. (1974), "Decision Analysis Expert Use," *Management Science*, 20, 1233–1241.
——— (1977), "Combining Expert Judgements: A Bayesian Approach," *Management Science*, 23, 679–693.
MURPHY, A.H., and EPSTEIN, E.S. (1967), "Verification of Probabilistic Predictions: A Brief Review," *Journal of Applied Meteorology*, 6, 748–755.
MURPHY, A.H., and WINKLER, R.L. (1977), "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Journal of the Royal Statistical Society*, Ser. C, 26, 41–47.
PRATT, J.W. (1962), "Must Subjective Probabilities be Realized as Relative Frequencies?" Unpublished seminar paper, Harvard University Graduate School of Business Administration.
ROBERTS, H.V. (1968), "On the Meaning of the Probability of Rain," Paper presented to First National Conference on Statistical Meteorology.
SHIMONY, A. (1955), "Coherence and the Axioms of Confirmation," *The Journal of Symbolic Logic*, 20, 1–28.
TITTERINGTON, D.M., MURRAY, G.D., MURRAY, L.S., SPIEGELHALTER, D.J., SKENE, A.M.M., HABBEMA, J.D.F., and GELPKER, G.J. (1981), "Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients (with Discussion)," *Journal of the Royal Statistical Society*, Ser. A, 144, 145–175.

Comment

JOSEPH B. KADANE*

Dawid gives a very interesting theorem to the effect that a coherent Bayesian feels almost certain he is well calibrated under conditions of feedback. It is an extension of Pratt's (1962) unpublished theorem on calibration. I have no criticism of the theorem; my comments concern Dawid's interpretation of it.

Coherence is a very mild set of constraints on a person's beliefs. It says roughly that those beliefs must be internally consistent, in the way described in Section 6 of Dawid's article. It does not imply that others do or should agree with a person who has coherent beliefs, nor that those subjective beliefs model well the predicted events. The person who is sure it will rain in the future on all odd numbered days and sure it will not rain on all even numbered days in a particular place is coherent. Yet, given my beliefs about the weather, I do not expect many days to pass before such a person is confronted with an event or subjective probability zero.

With this as background, let us reconsider the meaning of the sentences, "We denote by $\mathcal{B}_i$ the totality of events known to the forecaster on day i; thus $\mathcal{B}_0 \subseteq \mathcal{B}_1 \subseteq \ldots$

The forecaster has an arbitrary subjective probability distribution π defined over $\mathcal{B}_\infty = U_{i=0}^{\infty} \mathcal{B}_i$." In order to elicit π, I must first anticipate for each day i in the future all the possible events that might occur and might influence my probability for precipitation on day i. This will surely include the results for days $1, \ldots, i-1$ (and my probabilities of precipitation on days $1, \ldots, i-1$) and may include data from other places, and new meterological theories that may have been made known to me on day i, and so on. Merely enumerating the elements of $\mathcal{B}_i$ is a job beyond human capability.

The assumption $\mathcal{B}_0 \subseteq \mathcal{B}_1 \subseteq \ldots$ says essentially that my memory is perfect, that I never forget an event that might be relevant. While mathematically easy to state, this is not trivial to accomplish.

Having enumerated the elements of $\mathcal{B}_i$ and remembered all past elements $\mathcal{B}_{i-1}$ so that $\mathcal{B}_{i-1} \subseteq \mathcal{B}_i$, I must now state, for each possibility in the set $\mathcal{B}_i$, what my precipitation probability would be were that the event to be observed. Furthermore, I must do this in a way that respects everything I said about $\mathcal{B}_{i-1}$. This is again an extremely difficult task, and one I am sure to want to approximate in practice. Fully to elicit π is to anticipate the possibility of all future new data and new discoveries, to anticipate when they will be published, and then to state how influential such data and discoveries would be to me. Such an elicitation is beyond human possibility as a practical matter.

Nonetheless, let us join Dawid in supposing such a distribution. In this case Dawid shows that I believe that ultimately the ξ-weighted proportion of events occurring will approach my ξ-weighted probability, provided only that the sum of the weights go to infinity. I do not find this unreasonable. It says that in the infinitely far future I believe I will learn everything (down to an irreducible stochastic nub) about whether it will rain tomorrow.

What finite sequence of events should persuade me that miscalibration is in fact occurring? Professor Dawid is vague on this point. In principle, no finite initial sequence constrains a limit in any way. Furthermore the hypothetical elicitation of π has already required me to state how I would respond to each element of $\mathcal{B}_i$. So why should I change anything now? Only if I have done a

* Joseph B. Kadane is Professor, Departments of Statistics and of Social Science, Carnegie-Mellon University, Pittsburgh, PA 15213.

© Journal of the American Statistical Association
September 1982, Volume 77, Number 379
Theory and Methods Section