The Comparison and Evaluation of Forecasters

Author(s): Morris H. DeGroot and Stephen E. Fienberg

Source: *Journal of the Royal Statistical Society. Series D (The Statistician)*, Mar. - Jun., 1983, Vol. 32, No. 1/2, Proceedings of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics (Mar. - Jun., 1983), pp. 12-22

Published by: Wiley for the Royal Statistical Society

Stable URL: http://www.jstor.com/stable/2987588

The Statistician 32 (1983)

The Comparison and Evaluation of Forecasters†

MORRIS H. DeGROOT and STEPHEN E. FIENBERG

Department of Statistics, Carnegie–Mellon University, Pittsburgh, PA 15213, USA

Abstract: In this paper we present methods for comparing and evaluating forecasters whose predictions are presented as their subjective probability distributions of various random variables that will be observed in the future, e.g. weather forecasters who each day must specify their own probabilities that it will rain in a particular location. We begin by reviewing the concepts of calibration and refinement, and describing the relationship between this notion of refinement and the notion of sufficiency in the comparison of statistical experiments. We also consider the question of interrelationships among forecasters and discuss methods by which an observer should combine the predictions from two or more different forecasters. Then we turn our attention to the concept of a proper scoring rule for evaluating forecasters, relating it to the concepts of calibration and refinement. Finally, we discuss conditions under which one forecaster can exploit the predictions of another forecaster to obtain a better score.

1 Introduction

In this paper we describe some concepts and methods appropriate for evaluating and comparing forecasters who repeatedly present their predictions of whether or not various events will occur in terms of their subjective probabilities of those events. The ideas we describe here are relevant in almost any situation in which forecasters must repeatedly make such probabilistic predictions, regardless of the particular subject matter or substantive area of the events being forecast. The forecaster might be an economist who at the beginning of each quarterly period makes predictions about unemployment, the rate of inflation, or Gross National Product in that quarter based on the values of various economic indicators; the forecaster might even make predictions using a large-scale econometric model of the United States economy based on hundreds of variables and econometric relations.

In a different field, the forecaster might be the weatherman for a television station who at the beginning of each day must announce his probability that it will rain during the day. For ease of exposition, we present our discussions here in the context of such a weather forecaster who day after day must specify his subjective probability x that there will be at least a certain amount of rain at some given location during a specified time interval of the day. We refer to the occurrence of this well-specified event simply as "rain". Thus, at the beginning of each day the forecaster must specify his probability of rain and at the end of each day he observes whether or not rain actually occurred.

The probability x specified by the forecaster on any particular day is called his prediction for that day. We shall make the realistic, and simultaneously simplifying, assumption that the prediction x is restricted to a given finite set of values $0=x_0<x_1<\ldots<x_k=1$. In many

† Presented at the Institute of Statisticians International Conference on Practical Bayesian Statistics, Cambridge, England, 21–24 July, 1982. This research was supported in part by the National Science Foundation under grant SES-7906386 and by the Office of Naval Research under contract N00014-80-C-0637.

problems of weather forecasting, these values are 0, 1/10, 2/10, 3/10, . . ., 1. Let X denote the set of allowable values, $\{x_0, x_1, \ldots, x_k\}$.

We assume that the forecaster's predictions can be observed over a large number of days, and we let $\nu(x)$ denote the frequency function or probability function of his predictions over these days. In order to ensure that the frequency function $\nu(x)$ is of practical value, it is necessary to assume that weather conditions remain stationary over the period of n days being studied. More precisely, in our model we assume that, at the end of each day j, the forecaster can observe the value of a vector M_j of relevant meteorological variables, one component of which is the indicator variable θ_j, which is 1 or 0 according to whether or not it rained on day j. If we consider a potentially infinite sequence of consecutive days, rather than just the finite number n, then we assume that the joint distribution of the sequence of random vectors $M_1, M_2 \ldots$, is stationary *a priori* in the usual sense of time series analysis or stochastic processes. That is, many days in advance, a forecaster's joint distribution of any sequence of k random vectors $M_{t_1}, M_{t_2}, \ldots, M_{t_k}$ is the same as that of $M_{t+t_1}, M_{t+t_2}, \ldots, M_{t+t_k}$, for all values of t and all values of k. We say "many days in advance", because a weather forecaster clearly may have a different distribution for tomorrow's vector M_1 than for one associated with a day a year from now, i.e. M_{366}. The purpose of this somewhat loose stationarity requirement is to assure the "comparability" of the events whose outcome is being predicted and the existence of a long-run frequency or limiting probability function $\nu(x)$.

The function $\nu(x)$ can be thought of either as the probability that the forecaster's prediction on a randomly chosen day will be x or as the proportion of days on which his prediction is x. Thus, $\nu(x)$ satisfies the properties of probability: $\nu(x) \geqslant 0$ for $x \in X$ and $\Sigma_x \nu(x) = 1$.

In section 2 we study the concepts of calibration and refinement for forecasters. The concept of calibration pertains to the agreement between a forecaster's predictions and the actual observed relative frequency of rain. Roughly speaking, a forecaster is said to be well calibrated if among those days for which his prediction is x, the long-run relative frequency of rain is also x. The concept of refinement pertains to how spread out or how sharp a forecaster's predictions are. Roughly speaking, the more concentrated the probability function $\nu(x)$ is near the values $x=0$ and $x=1$, the more refined the forecaster is. These concepts form the basis for comparing different forecasters.

In section 3 we relate the concepts of calibration and refinement to the concept of sufficiency in the theory of the comparison of statistical experiments. Different notions of sufficiency in statistical inference are discussed in this context.

In section 4 we review the use of the Brier or quadratic scoring rule for evaluating a forecaster, showing that this scoring rule can be naturally partitioned into components separately related to calibration and refinement. We then study the question of when an observer can use a forecaster's predictions to obtain a better score than the forecaster himself, and show that such an improvement can be achieved by the observer essentially if and only if the forecaster is not well-calibrated.

Finally, in section 5, we present results similar to those just described for other special scoring rules and, more generally, for any arbitrary strictly proper scoring rule.

2 Calibration and refinement

For any particular weather forecaster, we let $\rho(x)$ denote the relative frequency, or conditional probability, of rain among all those days on which the forecaster's prediction was x. If the forecaster's probability x is to be directly interpretable in terms of the actual chances of rain, then $\rho(x)$ should be close to x. In accordance with this notion, a forecaster is said to be *well calibrated* if $\rho(x)=x$ for every value of x such that $\nu(x)>0$ (see, e.g., Dawid, 1982). In meteorology, the concept of calibration is called validity (Miller, 1962) or reliability (Murphy, 1973), and a well-calibrated forecaster is called *perfectly reliable*.

Pratt (1962) and Dawid (1982) have developed theorems which show that if a forecaster has a joint probability distribution for all the random meteorological vectors $M_1, M_2, \ldots$, that he will observe over an infinite sequence of days, and if his prediction each day is the appropriate conditional probability of rain given all the past observations, then according to his joint probability distribution, there is probability 1 that he must be well calibrated over the infinite sequence of days. In effect, these results (see also Lad, 1982) show that simply being coherent in the sense of de Finetti (1937) is tantamount to assigning subjective probability 1 to the prospect of being well calibrated over the infinite sequence.

In practice, however, there are two reasons why a forecaster may not be well calibrated. First, his predictions can be observed for only a finite number of days. Second, and more importantly, there is no inherent reason why his predictions should bear any relation whatsoever to the actual occurrence of rain.

Although being well calibrated is widely regarded as a desirable characteristic of a forecaster, a forecaster can usually make himself well calibrated by keeping track of the evolving values of $\rho(x)$ day after day and lying, i.e. specifying predictions that do not represent his honest subjective probabilities, in order to adjust certain values of $\rho(x)$ as necessary (DeGroot, 1979). Furthermore, as Murphy and Winkler (1977) and Dawid (1981) have pointed out, even if a forecaster's honest subjective probabilities make him well calibrated, his predictions are not necessarily accurate in all respects and they are not necessarily of much use to anyone.

For example, suppose that it is known that the overall relative frequency of rain is μ. In meteorology, μ is sometimes called the climatological probability. Then a forecaster whose prediction is μ on each day will be calibrated in the long run, but his predictions are obviously useless. At the other extreme, a forecaster whose prediction each day is either $x=0$ or $x=1$, and who is always correct, is also well calibrated and displays perfect foresight.

In order to compare various well-calibrated forecasters, we introduce a concept of refinement among them. Consider two well-calibrated forecasters A and B, and suppose that we know the functions $\nu_A(x)$ and $\nu_B(x)$ that characterize their predictions. Since both forecasters are well calibrated, $\rho(x)=x$ for each of them. Then forecaster A is said to be *at least as refined* as forecaster B if, from A's prediction on each day and an auxiliary randomization based only on a table of random numbers, we can simulate a prediction with the same stochastic properties as B's predictions. In other words, A is at least as refined as B if we can artificially generate a well-calibrated forecaster with the same probability function $\nu_B(x)$ as B simply by passing A's predictions through a noisy channel. The mathematical definition is as follows.

A *stochastic transformation* $h(x|y)$ is a function defined for all $x\in X$ and $y\in X$ such that

$$h(x|y)\geqslant 0 \qquad \text{for } x\in X \text{ and } y\in X$$

$$\sum_{x\in X} h(x|y)=1 \qquad \text{for } y\in X \tag{2.1}$$

Forecaster A is at least as refined as forecaster B if there exists a stochastic transformation h such that the following relations are satisfied:

$$\sum_{y\in X} h(x|y)\,\nu_A(y)=\nu_B(x) \qquad \text{for } x\in X \tag{2.2}$$

$$\sum_{y\in X} h(x|y)\,y\nu_A(y)=x\nu_B(x) \qquad \text{for } x\in X \tag{2.3}$$

The function h determines the auxiliary randomization that is to be carried out. If A makes the prediction y on a particular day, then we generate a prediction x by means of an auxiliary randomization in accordance with the conditional probability distribution $h(x|y)$. The relation (2.2) guarantees that in this way we will make each prediction x with the same

frequency $\nu_B(x)$ that B does, and the relation (2.3) guarantees that our predictions will again be well calibrated.

In these terms, the forecaster who makes the same prediction μ each day is *least-refined* in the sense that any other well-calibrated forecaster is at least as refined as he is. It is possible that μ is not one of the allowable predictions $x_0, x_1, \ldots, x_k$. In that case, there is a value of i ($i=0, 1, \ldots, k-1$) such that $x_i < \mu < x_{i+1}$, and we have shown in DeGroot and Fienberg (1982) that a forecaster who uses only the values x_i and x_{i+1} as his predictions in such a way that he is well calibrated will now be least-refined. At the other extreme, the forecaster whose prediction each day is either $x=0$ or $x=1$ and who is always correct is *most-refined* in the sense that he is at least as refined as any other well-calibrated forecaster.

When two well-calibrated forecasters are being compared, it is quite possible that neither one of them is at least as refined as the other. The two forecasters simply might not be comparable in terms of refinement, and it can be difficult to establish this fact. We might very well find ourselves unable to construct a stochastic transformation h that satisfies (2.2) and (2.3), but also unable to convince ourselves that such a transformation does not exist. Or we might feel fairly certain that forecaster A is at least as refined as forecaster B, but again find ourselves unable to construct a specific h that satisfies (2.2) and (2.3). The following result, which was originally presented and proven in DeGroot and Fienberg (1982), is useful in that it provides a simple necessary and sufficient condition that eliminates any need for being able to construct an appropriate stochastic transformation.

Theorem 1
Consider two well-calibrated forecasters A and B. Then A is at least as refined as B if and only if the following inequalities are satisfied:

$$\sum_{i=0}^{j-1} (x_j - x_i)\{\nu_A(x_i) - \nu_B(x_i)\} \geq 0 \qquad \text{for } j=1, \ldots, k-1 \tag{2.4}$$

As a simple example, if A and B are well-calibrated forecasters with the following probability functions, then neither A nor B is at least as refined as the other:

$$\nu_A(x) = \begin{cases} 0{\cdot}1 & \text{for } x=0 \\ 0{\cdot}8 & \text{for } x=0{\cdot}5 \\ 0{\cdot}1 & \text{for } x=1 \end{cases} \tag{2.5}$$

$$\nu_B(x) = \begin{cases} 0{\cdot}5 & \text{for } x=0{\cdot}1 \\ 0{\cdot}5 & \text{for } x=0{\cdot}9 \end{cases} \tag{2.6}$$

In summary we note that if one well-calibrated forecaster A is at least as refined as another well-calibrated forecaster B, and if we must choose between learning the prediction of A or learning the prediction of B, then we should choose to learn the prediction of A regardless of the purposes for which we will use the prediction.

3 Concepts of sufficiency

In this section, we consider the comparison of forecasters who are not necessarily well calibrated. We begin by extending the concepts that we have just described for well-calibrated forecasters to the class of all forecasters.

Consider two arbitrary forecasters A and B, and suppose that we know the functions $\nu_A(x)$ and $\rho_A(x)$ that characterize the predictions of A and we know the functions $\nu_B(x)$ and $\rho_B(x)$ that characterize B. Then forecaster A is said to be *sufficient* for forecaster B if, from A's prediction on each day and an auxiliary randomization based only on a table of random numbers, we can simulate a prediction with the same stochastic properties as B's predictions.

In other words, A is sufficient for B if we can artificially generate a forecaster with the same functions $\nu_B(x)$ and $\rho_B(x)$ as B simply by passing A's predictions through a noisy channel. The mathematical definition is as follows: forecaster A is *sufficient* for forecaster B if there exists a stochastic transformation h such that the following relations are satisfied:

$$\sum_{y \in X} h(x|y)\, \nu_A(y) = \nu_B(x) \qquad \text{for } x \in X \tag{3.1}$$

$$\sum_{y \in X} h(x|y)\, \rho_A(y)\, \nu_A(y) = \rho_B(x)\, \nu_B(x) \qquad \text{for } x \in X \tag{3.2}$$

When the forecasters A and B are well calibrated, this relationship of sufficiency reduces to the relationship of refinement discussed in section 2. It can be seen that (3.1) is the same as (2.2) and, when A and B are well calibrated, (3.2) is the same as (2.3).

Up to this point, we have characterized the predictions of a forecaster in terms of the two functions $\nu(x)$ and $\rho(x)$, but there is an alternative and equivalent characterization that permits our discussion to be related more directly to standard statistical theory (see, e.g., Lindley *et al.* 1979; Lindley 1981). Let θ denote the indicator of rain, so $\theta = 1$ if rain occurs on a particular day and $\theta = 0$ otherwise. For any given forecaster let $f(x|\theta)$ denote the conditional probability function of the forecaster's predictions given θ. Thus, $f(x|1)$ represents the frequency function of the forecaster's predictions on days when rain actually occurs, and $f(x|0)$ represents the frequency function on days when rain does not occur. It follows that for $x \in X$,

$$f(x|1) = \rho(x)\, \nu(x)/\mu \tag{3.3}$$

$$f(x|0) = \{1 - \rho(x)\}\, \nu(x)/(1-\mu) \tag{3.4}$$

From (3.3) and (3.4) we see that knowledge of the functions $\nu(x)$ and $\rho(x)$ is equivalent to knowledge of the functions $f(x|1)$ and $f(x|0)$. Hence, the two functions $f(x|1)$ and $f(x|0)$ characterize the forecaster's predictive behaviour. In terms of this characterization we can now apply the original definition of sufficiency in the theory of the comparison of statistical experiments as given by Blackwell (1951, 1953) (see also Blackwell and Girshick, 1954, Chapter 12; or DeGroot, 1970, sec. 14.17). To apply this theory, we simply regard learning the prediction of a forecaster as observing the outcome of a particular statistical experiment. Then forecaster A is sufficient for forecaster B if there exists a stochastic transformation h such that the following relation is satisfied:

$$\sum_{y \in X} h(x|y) f_A(y|\theta) = f_B(x|\theta) \qquad \text{for } x \in X \text{ and } \theta = 0, 1 \tag{3.5}$$

The following theorem shows that this definition of sufficiency is equivalent to the definition given at the beginning of this section. The proof is direct and is given in DeGroot and Fienberg (1982).

Theorem 2
A stochastic transformation h satisfies (3.5) if and only if it satisfies (3.1) and (3.2).

As before, not all forecasters are comparable in terms of the relationship of sufficiency; it induces only a partial ordering among all forecasters. It is not necessarily true, however, that if forecaster A is sufficient for forecaster B then A is at least as good a forecaster as B. For example, suppose that A never makes a prediction other than $x = 0$ or $x = 1$, but that he is always wrong about whether or not it is going to rain. Then A is sufficient for every other forecaster, even though he is the worst possible forecaster. If we know that A is always wrong, his predictions are just as useful to us as those of a forecaster who is always correct.

If forecaster A is sufficient for forecaster B, and if we must choose between learning the prediction of A or learning the prediction of B, then we should choose to learn the prediction

of A regardless of the purposes for which we will use the prediction. It is important to emphasize that this concept of sufficiency is based on our having to choose between learning the prediction of A or the prediction of B. Even though A is sufficient for B in this sense, it is quite possible that by learning the prediction of B in addition to the prediction of A, we might gain more information than we gain from the prediction of A alone.

As we have seen from (3.5), in order to determine whether A is sufficient for B, we need only know the marginal probability functions $f_A(x|\theta)$ and $f_B(x|\theta)$ of A and B separately. However, in order to determine whether forecaster A is sufficient for both himself and forecaster B together, we must know the joint probability function of the predictions of A and B on any given day. By analogy with the notation already developed, let $f(x, y|1)$ denote the relative frequency of days on which the prediction of A is x and the prediction of B is y among all days on which $\theta=1$. Similarly, let $f(x, y|0)$ denote this relative frequency among all days on which $\theta=0$.

Let $h(y|x, 1)$ and $h(y|x, 0)$ denote the conditional probability functions of y given x, derived from the joint probability function $f(x, y|1)$ and $f(x, y|0)$ respectively. In accordance with the classical theory of sufficient statistics, we say that forecaster A is *sufficient for the pair* of forecasters A and B, or simply that A is sufficient for (A, B), if the following relation is satisfied:

$$h(y|x, 1)=h(y|x, 0) \qquad \text{for } x\in X \text{ and } y\in X \tag{3.6}$$

If (3.6) is satisfied, then the common conditional probability function h can be used as the stochastic transformation in (3.5) and will satisfy that relation. Thus, if A is sufficient for (A, B), then A is sufficient for B. The converse, however, is not necessarily true. In particular, since the relation (3.5) involves only the functions $f_A(x|\theta)$ and $f_B(x|\theta)$ it is possible for A to be sufficient for B and for the joint probability function $f(x, y|\theta)$ to be given by the product of the marginal probability functions. We begin with any pair of forecasters A and B such that A is sufficient for B. Then we simply multiply the functions $f_A(x|\theta)$ and $f_B(x|\theta)$ together to yield

$$f(x, y|\theta)=f_A(x|\theta)f_B(y|\theta) \tag{3.7}$$

In this case, the predictions of A and B are conditionally independent given θ. Thus, learning the prediction of B will typically add new information about θ beyond that contained in the prediction of A, and it follows that A will not be sufficient for (A, B).

We conclude this section with another characterization of problems in which forecaster A is sufficient for the pair (A, B), this time in terms of the notation of section 2. Let $\nu(x, y)$ denote the joint probability function of the predictions x and y, i.e. $\nu(x, y)$ is the relative frequency of days on which the prediction of A is x and the prediction of B is y. Furthermore, let $\rho(x, y)$ denote the conditional probability of rain among those days on which the predictions are x and y. Then

$$h(y|x, 1)=\frac{\rho(x, y)\,\nu(x, y)}{\rho_A(x)\,\nu_A(x)} \tag{3.8}$$

$$h(y|x, 0)=\frac{\{1-\rho(x, y)\}\,\nu(x, y)}{\{1-\rho_A(x)\}\,\nu_A(x)} \tag{3.9}$$

It follows from (3.8) and 3.9) that expression (3.6) will be satisfied if and only if $\rho(x, y)=\rho_A(x)$ for all $x\in X$ and $y\in X$. Thus, we have established the following result:

Theorem 3

Forecaster A is sufficient for the pair of forecasters (A, B) if and only if

$$\rho(x, y)=\rho_A(x) \qquad \text{for } x\in X \text{ and } y\in X \tag{3.10}$$

4 The Brier or quadratic scoring rule

It has often been suggested in the statistical literature that a forecaster's predictions over a sequence of days can be evaluated by the use of a *scoring rule* which assigns a numerical value, or score, each day based on the forecaster's prediction x and the observation of whether or not rain occurred, i.e. the observation of θ. One of the earliest scoring rules proposed for meteorological forecasts is the quadratic scoring rule $(x-\theta)^2$, introduced by Brier (1950). With this rule the forecaster's score on any given day reduces to the square of the probability that he assigned to the event, either rain or no rain, that did *not* occur.

The forecaster's overall Brier Score (BS) is the average of the values of $(x-\theta)^2$ over all of the days on which predictions are made. When forecasters are evaluated by this scoring rule, the forecaster with the smallest BS receives the highest evaluation. Among those days on which a forecaster's prediction is x, his score will be $(x-1)^2$ with relative frequency $\rho(x)$ and it will be x^2 with relative frequency $1-\rho(x)$. Since the relative frequency with which he makes the prediction x is $\nu(x)$, we find that

$$\mathrm{BS}=\sum_{x\in X}\nu(x)\,[\rho(x)\,(x-1)^2+\{1-\rho(x)\}x^2] \tag{4.1}$$

After some algebra, this can be expressed in the form

$$\mathrm{BS}=\sum_{x\in X}\nu(x)\,\{x-\rho(x)\}^2+\sum_{x\in X}\nu(x)\,\rho(x)\,\{1-\rho(x)\} \tag{4.2}$$

The first summation on the right-hand side of (4.2) is a measure of the calibration of the forecaster. The closer $\rho(x)$ is to x the smaller this component will be. If the forecaster is well calibrated, this component is zero. The second summation on the right-hand side of (4.2) is a measure of the refinement or sufficiency of the forecaster. The more concentrated the values of $\rho(x)$ are near 0 and 1 the smaller this component will be. We show in DeGroot and Fienberg (1982) that, if forecaster A is sufficient for forecaster B, then this second component will be at least as large for B as it is for A. Tukey *et al.* (1965) have suggested a variant of BS in which the two summations on the right-hand side of (4.2) are given different weights. Murphy (1972) has given a related partition of BS.

Suppose now that we know the functions $\nu(x)$ and $\rho(x)$ that characterize a particular forecaster's predictions. Is it possible for us to use his predictions, and no other relevant meteorological information, to make our own predictions and to attain a smaller value of BS than the forecaster himself?

The forecaster's BS is given by expression (4.1). In order for us to make our predictions, we must choose a stochastic transformation $h(x|y)$ to be used in the following manner. If the forecaster's prediction on a given day is y, then we choose our prediction at random from X in accordance with the conditional distribution $h(x|y)$. With this procedure, our predictions are characterized by the functions

$$\nu_0(x)=\sum_{y\in X}h(x|y)\,\nu(y) \tag{4.3}$$

$$\rho_0(x)=\sum_{y\in X}h(x|y)\,\rho(y)\,\nu(y)/\nu_0(x) \tag{4.4}$$

It follows from expressions (4.1), (4.3) and (4.4), after some algebra, that our Brier Score BS_0 is given by

$$\mathrm{BS}_0=\sum_{y\in X}\nu(y)\sum_{x\in X}\{\rho(y)\,(x-1)^2+\{1-\rho(y)\}\,x^2]\,h(x|y) \tag{4.5}$$

For each fixed value of y, the summation over x in expression (4.4) yields a weighted average of the values

$$\rho(y)\,(x-1)^2+\{1-\rho(y)\}\,x^2 \tag{4.6}$$

with weights given by the conditional probabilities $h(x|y)$. Hence, in order to minimize BS_0, we should choose the conditional distribution $h(x|y)$ to put all the probability on the value of x that minimizes expression (4.6). If $\rho(y)$ lies in the allowable set of predictions X then all is well and the value that minimizes expression (4.6) is $x=\rho(y)$, i.e. we make the forecaster well calibrated. If $\rho(y)$ does not lie in the allowable set X then we come as close to the minimum of expression (4.6) as possible, by setting x equal to the permissible value closest to $\rho(y)$, i.e. we make the forecaster almost well calibrated.

Thus, when the forecaster's prediction is y, we should make the prediction $\rho(y)$ (ignoring the minor detail that $\rho(y)$ might not lie in the allowable set of predictions X). If the forecaster is well calibrated then we will simply repeat his prediction and attain the same Brier Score that he does. If he is not well calibrated, however, then by making the prediction $\rho(y)$ when his prediction is y we can attain a smaller Brier Score than he does. Even if $\rho(y)$ is not in X for some predictions $y\in Y\subset X$, we can still attain a smaller Brier Score than he does by choosing the value in X closest to $\rho(y)$ whenever $y\in Y$, and choosing $\rho(y)$ otherwise, unless the forecaster's value y is already as close as possible to $\rho(y)$ for all $y\in Y$. In summary, we can improve upon the forecaster's BS if and only if he is not well calibrated, and has not tried to make himself so.

Now suppose that we know the functions $\nu(x, y)$ and $\rho(x, y)$ that characterize jointly the predictions x and y of two forecasters A and B. How can we use the two predictions x and y, and no other meteorological information, to minimize our Brier Score? By an argument analogous to the one just presented, we can show that our prediction on any given day should be $\rho(x, y)$. In this way we can attain a Brier Score that will be no larger than the score of either of the other forecasters and will typically be smaller than both. If forecaster A is sufficient for the pair (A, B), then by Theorem 3, our prediction will simply be $\rho_A(x)$ and it is not helpful to us to learn the prediction of forecaster B. Again, if $\rho(x, y)$ is not in X, we need to use the value in X closest to it.

5 Strictly proper scoring rules

The Brier or quadratic scoring rule introduced in section 4 has the property that the forecaster will minimize his subjective expected score on any particular day by stating as his prediction x the value that is his actual subjective probability p of rain for that day. In symbols, his expected score when he makes the prediction x is

$$p(x-1)^2+(1-p)x^2 \tag{5.1}$$

and this is minimized uniquely at the point $x=p$. In this section we consider other scoring rules that have this property.

To bring our discussion in this section into agreement with much of the statistical literature, we shall consider scoring rules with the property that the forecaster desires to *maximize* his score rather than minimize it. Obviously, any scoring rule that is to be maximized is the negative of one to be minimized, so the development can be given equally effectively in either setting. In particular, the negative of the Brier Score is one that should be maximized.

An arbitrary scoring rule has the following form: suppose that the forecaster's prediction is x. Then, if rain occurs the forecaster receives a score $g_1(x)$ and if rain does not occur he receives a score $g_2(x)$. Since we assume that the forecaster desires to maximize his score, it is reasonable to assume that $g_1(x)$ is an increasing function of x and that $g_2(x)$ is a decreasing function of x.

If the forecaster's actual subjective probability of rain on a particular day is p and he makes the prediction x, then his expected score is

$$pg_1(x)+(1-p)g_2(x) \tag{5.2}$$

A *proper scoring rule* is one for which expression (5.2) is maximized when $x=p$. A *strictly proper scoring rule* is one for which $x=p$ is the *only* value of x that maximizes expression (5.2). An interesting discussion of these rules, with historical references, is given by Stael von Holstein (1970, sec, 3.2).

The negative of the Brier Score is a strictly proper scoring rule with $g_1(x)=-(x-1)^2$ and $g_2(x)=-x^2$. Another interesting strictly proper scoring rule is the logarithmic rule (Good, 1952) in which $g_1(x)=\log x$ and $g_2(x)=\log(1-x)$. Both the Brier and the logarithmic rules have the symmetry property that $g_1(x)=g_2(1-x)$.

Various ways of characterizing strictly proper scoring rules have been presented in the literature. If $\alpha(t)$ is a positive continuous function on the interval $0\leqslant t\leqslant 1$ for which the integral to be given in (5.3) exists, then the scoring rule defined as follows is *strictly proper* (see Shuford *et al.*, 1966: or Stael von Holstein, 1970):

$$g_1(x)=\int_0^x \alpha(t)\,dt, \qquad g_2(x)=\int_x^1 \frac{t}{1-t}\,\alpha(t)\,dt \tag{5.3}$$

If $J(x)$ is a strictly convex, differentiable function on the interval $0\leqslant x\leqslant 1$, then the following scoring rule is also strictly proper (see Savage, 1971):

$$g_1(x)=J(x)+(1-x)\frac{dJ(x)}{dx}, \qquad g_2(x)=J(x)-x\frac{dJ(x)}{dx} \tag{5.4}$$

Both expressions (5.3) and (5.4) characterize the class of strictly proper scoring rules under mild regularity conditions on g_1 and g_2.

The theory underlying the use of strictly proper scoring rules is that they encourage the forecaster to give his actual subjective probability of rain as his prediction, because by so doing he maximizes his expected score. The shortcoming of this theory is that in general there is no reason to believe that the forecaster is trying to maximize the expected value of his score rather than the expected value of some other function of his score. In other words, there is no reason to believe that the forecaster's utility function is simply a linear function of his score rather than some increasing but non-linear function.

This shortcoming becomes more pronounced when we consider the forecaster's overall score as the average of his scores over the days on which predictions are made. The forecaster's objective may be to maximize the probability that his overall score will be higher than that of a rival forecaster. If the forecaster is actually being paid on the basis of his score, he may place special importance on trying to ensure that his payments do not fall below some fixed level. Considerations like these may well lead the forecaster on occasion to make a prediction that does not represent his actual subjective probability of rain. On the other hand, if the forecaster's utility function for payments were known, then a schedule of payments could presumably be worked out that would encourage him to want to maximize his expected overall score. Comments along these same lines are given by Stael von Holstein (1970).

We conclude this section by noting that every strictly proper scoring rule can be partitioned in a manner analogous to that given in expression (4.2) for the Brier Score.

Theorem 4

If a forecaster's predictions are characterized by the functions $\nu(x)$ and $\rho(x)$, and if a proper scoring rule is specified by the functions $g_1(x)$ and $g_2(x)$, then the forecaster's overall score S can be expressed in the form $S=S_1+S_2$, where

$$S_1=\sum_{x\in X}\nu(x)\,[\![\rho(x)\,[g_1(x)-g_1\{\rho(x)\}]+\{1-\rho(x)\}\,[g_2(x)-g_2\{\rho(x)\}]]\!] \tag{5.5}$$

$$S_2=\sum_{x\in X}\nu(x)\,\phi\{\rho(x)\} \tag{5.6}$$

and

$$\phi(t) = tg_1(t) + (1-t)\, g_2(t) \qquad \text{for } 0 \leqslant t \leqslant 1 \tag{5.7}$$

If the scoring rule is strictly proper then $\phi(t)$ is strictly convex and S_1 attains its maximum value only when $\rho(x) = x$ for every value of x such that $\nu(x) > 0$.

According to Theorem 4, whose proof will be given elsewhere, if the scoring rule is strictly proper, then we can regard S_1 as a measure of the forecaster's calibration. It is zero only for a well-calibrated forecaster and it is negative otherwise. Furthermore, we can regard S_2 as a measure of the refinement or sufficiency of the forecaster. Since ϕ is a convex function of $\rho(x)$, we can show that, if two forecasters A and B are both well calibrated and A is at least as refined as B, then the value of S_2 will be at least as large for A as it is for B (see DeGroot and Fienberg (1982) for a related discussion).

In general, since $\phi(t)$ is a strictly convex function on the interval $0 \leqslant t \leqslant 1$, it attains its maximum value at one endpoint or the other. Hence, the more concentrated the values of $\rho(x)$ are near this endpoint, the larger the value of S_2 will be. If $g_2(x) = g_2(1-x)$, then $\phi(t)$ will be symmetric with respect to $t = \frac{1}{2}$. In this case, the more concentrated the values of $\rho(x)$ are near 0 and 1, the larger the value of S_2 will be. As mentioned earlier, both the quadratic and the logarithmic scoring rules have this symmetry.

6 Data arrays for comparison

The results described in this paper suggest how we should actually summarize data for the evaluation and comparison of forecasters, no matter which scoring rule we ultimately wish to use. For convenience, we restrict attention here to two forecasters, A and B, but our remarks generalize with little trouble.

All of the information needed to evaluate and compare a pair of forecasters making predictions over the *same set* of n days, can be summarized in the form of a $(k+1) \times (k+1) \times 2$ table of counts. Dimensions 1 and 2 of this table correspond to the $k+1$ possible values of the predictions of forecasters A and B, respectively. Dimension 3 corresponds to the two outcomes, rain ($\theta = 1$) and no rain ($\theta = 0$). Thus the frequency f gives the proportion of the n days on which A predicts x_i, B predicts x_j, and the outcome is θ. Note that we do not make any use of the other meteorological variables discussed in section 1.

To check on the calibration of forecasters A and B we need only compare the relative frequencies of rain from each of the two-way marginals, i.e. $\{f_{i+1}/f_{i++}\}$ and $\{f_{+j1}/f_{+j+}\}$ corresponding to $\{\rho_A(x_i)\}$ and $\{\rho_B(x_j)\}$, with the values of the predictions, i.e. $\{x_0, x_1, \ldots, x_k\}$. To determine if forecaster A is at least as refined as forecaster B we apply Theorem 1 of section 2, which requires only the one-way marginals $\{f_{i++}\}$ and $\{f_{+j+}\}$, corresponding to $\{\nu_A(x_i)\}$ and $\{\nu_B(x_j)\}$.

In section 3, we noted the relationship of calibration and refinement to the concept of sufficiency in the statistical theory of experiments. To demonstrate that forecaster A is sufficient for forecaster B we again need only compare the relative frequencies of predictions given rain and given no rain, i.e. $\{f_{i+\theta}/f_{++\theta}\}$ and $\{f_{+j\theta}/f_{++\theta}\}$, to see if a stochastic transformation h satisfying expression (3.5) exists. As we noted in section 2, searching for such an h even in the case that both forecasters are well calibrated may prove both frustrating and fruitless.

What we would like is a way to extract the stochastic transformation h from the data array $\{f_{ij\theta}\}$ itself. We can do this in the special case where A is sufficient for (A, B), for then B's predictions are *conditionally independent* of the outcomes given A's predictions (see Theorem 3). We can check for conditional independence directly in $\{f_{ij\theta}\}$. From expression (3.6) and the ensuing discussion, we see that when A is jointly sufficient for (A, B) the transformation h is given by the common conditional frequencies $\{f_{ij+}/f_{i++}\}$.

The fact that the relationships of interest for the comparison and evaluation of a pair of forecasters all seem to depend only on two-way marginal totals suggests a possible link with loglinear models and in particular the model of no second-order interaction (e.g. see Bishop *et al.*, 1975; Fienberg, 1980). Except for the direct interpretation of joint sufficiency in terms of conditional independence we have yet to discover such a link.

Finally, we note that for predictions on a finite number of days, calibration and other relationships discussed here hold only approximately due to sampling variability – a matter we have not addressed in this paper.

References

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, Berkeley.

Blackwell, D. (1953). Equivalent comparison of experiments. *Annals of Mathematical Statistics* **24,** 265–72.

Blackwell, D. and Girschick, M. A. (1954). *Theory of Games and Statistical Decisions*. John Wiley, New York.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78,** 1–3.

Dawid, A. P. (1981). Discussion of papers on improving judgements using feedback. In *Bayesian Statistics, Proceedings of the First International Meeting* (eds J. M. Bernardo *et al.*), pp. 418–19. University Press, Valencia, Spain.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association* **77,** 605–10.

de Finetti, B. (1937). Foresight: its logical laws, its subjective sources (English translation from French). In *Studies in Subjective Probability* (1964) (eds H. E. Kyburg and H. E. Smokler), pp. 93–158. John Wiley, New York.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

DeGroot, M. H. (1979). Comments on Lindley *et al. Journal of the Royal Statistical Society A* **142,** 172–3.

DeGroot, M. H. and Fienberg, S. E. (1982). Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics III*, Vol. 1 (eds S. S. Gupta and J. O. Berger), pp. 291–314. Academic Press, New York.

Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data*, 2nd edn. MIT Press, Cambridge, Mass.

Good, I. J. (1952) Rational decisions. *Journal Royal Statistical Society B*, **14,** 107–14.

Lad, F. (1982). The calibration question. Unpublished memorandum, Department of Economics, University of Utah.

Lindley, D. V. (1981). The improvement of probability judgements. Unpublished manuscript.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society A* **142,** 146–80.

Miller, R. G. (1962). Statistical prediction by discriminant analysis. *Meteorological Monographs* **4,** No. 25.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: part I. Two-state situation. *Journal of Applied Meteorology* **11,** 273–82.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology* **12,** 595–600.

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics* **26,** 41–7.

Pratt, J. W. (1962). Must subjective probabilities be realized as relative frequencies? Unpublished seminar paper. Harvard University Graduate School of Business Administration.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66,** 783–801.

Shuford, E. H., Jr, Albert, A. and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika* **31,** 125–45.

Stael von Holstein, C.-A. S. (1970). *Assessment and Evaluation of Subjective Probability Distributions*. Economic Research Institute, Stockholm School of Economics, Stockholm.

Tukey, J. W., Mosteller, F. and Fienberg, S. E. (1965). Scoring probability forecasts. Memorandum NS-37, Department of Statistics, Harvard University.